Estimation par Minimum de Contraste Régulier et Heuristique de Pente en Sélection de Modèles

Adrien Saumard

N° d'ordre : 4020 ANNÉE 2010

UNIVERSITÉ DE RENNES 1

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Mathématiques et Applications

École doctorale MATISSE

présentée par

Adrien Saumard

préparée à l'unité de recherche n°6625 CNRS - IRMAR
Institut de Recherche Mathématiques de Rennes
U.F.R. de Mathématiques

Intitulé de la thèse

Estimation par Minimum de Contraste Régulier et Heuristique de Pente en Sélection de Modèles

Thèse soutenue à Rennes
le 22 octobre 2010

devant le jury composé de :

Philippe BERTHET Université Toulouse III	Professeur Directeur de thèse
Lucien BIRGÉ Université Paris VI	Professeur Examinateur
Olivier CATONI CNRS et ENS	Directeur de Recherche Examinateur
Bernard DELYON Université Rennes 1	Professeur Examinateur
Pascal MASSART Université Paris XI, Orsay	Professeur Rapporteur
Jian-Feng YAO Université Rennes 1	Professeur Examinateur

Je dédie cette thèse à mon père, qui a su me transmettre
son goût pour la logique et le raisonnement.

What I cannot create, I do not understand.

RICHARD FEYNMAN, *pensée*[1]

[1]Cette pensée a été écrite par Richard Feynman en haut à gauche de son tableau de travail. Elle apparaît sur une photographie qu'a pris Robert Paz du bureau de Feynman au lendemain de sa mort en 1988. Cette photo est archivée au California Institute of Technology et disponible à l'adresse suivante : http://users.physik.fu-berlin.de/~kleinert/feynman/last16s.jpg.

Remerciements

Mes premiers remerciements vont à mon directeur de thèse, Philippe Berthet. Tout d'abord, je te remercie Philippe pour la qualité de ton enseignement en Master 2. De part la beauté des mathématiques que tu m'as fait découvrir et l'enthousiasme très communicatif avec lequel tu les exposais, il m'est très tôt apparu comme une évidence que tu étais la personne que je recherchais pour m'encadrer en thèse. Je me reconnaissais en effet parfaitement - et je me reconnais toujours - dans la démarche qui est tienne de rechercher des principes mathématiques fondamentaux en statistique, ou en d'autres termes de plaider pour un effort de généralisation dans ce domaine. Ainsi, tu as placé le processus empirique au centre de ton discours, comme objet qui sous-tend toute analyse statistique, et tu as insisté sur une certaine vision trajectorielle de cet objet - par approximation forte, i.e. par un couplage à un processus Gaussien en distance uniforme sur les trajectoires, par des lois de recouvrement (des trajectoires dans un espace convenable), etc... Dans les exemples traités en cours, la puissance de cette démarche était transparente, et je me considère comme chanceux d'avoir pu apprendre très tôt ces outils, qui sont méconnus actuellement des jeunes chercheurs en statistique, bien que la démarche ait déjà fait ses preuves sur $\mathbb{R}$, notamment via KMT. Mon sentiment est que, loin d'être obsolète, le point de vue que tu m'as enseigné pourrait bien revenir au coeur du débat en grande dimension, car on peut espérer qu'une généralisation convenable de KMT - problème complètement ouvert à l'heure actuelle - apporte des minorations suffisamment informatives aux fluctuations d'échantillonnage contrôlées par les inégalités de concentration de type Talagrand. Mes dernières préoccupations me laissent d'ailleurs à penser que cette piste est porteuse pour aborder le problème de la validité de l'heuristique de pente avec une grande collection de modèles...

Je te remercie enfin et surtout pour la grande gentillesse que tu as toujours eu à mon égard, et le respect profond que tu as de mes choix scientifiques. En effet, tu m'as toujours montré que tu croyais en moi, et tu m'as encouragé à suivre mes aspirations et mes propres intuitions, deux choses inestimables dans la vie d'un thésard. Je suis sûr que ta méthode et ton contact contribuent à drainer des esprits autonomes et passionnés vers les statistiques dites fondamentales.

Deux autres personnes ont joué un rôle prépondérant pendant l'écriture de mon mémoire. Tout d'abord Matthieu Lerasle, qui a relu en grande partie la première version de mon mémoire, et a été le premier à valider mes preuves. Je te remercie donc chaleureusement Matthieu, pour ta générosité, ta franchise et la qualité de tes conseils d'écriture. La deuxième personne qui mérite toute ma gratitude est Sylvain Arlot, qui a très généreusement accepté de relire en détails plus de quatre-vingts pages de la seconde version de mon mémoire. Tes conseils et remarques quant à la refonte des chapitres et l'explicitation des arguments dans les preuves pour une plus grande netteté, ont débouché directement sur la forme finale de mon mémoire. Je ne saurais que trop te remercier d'avoir su me prêter main forte au moment où je suis venu te demander de l'aide. Tant pour Matthieu que pour Sylvain, nos discussions ont mis plus précisement à jour des intérêts scientifiques communs et je souhaite que nos rapports s'épanouissent encore dans les collaborations que nous avons projetées.

Je tiens aussi à adresser mes sincères remerciements à Pascal Massart qui, non comptant de me faire le privilège d'être mon rapporteur, a su avec une bienveillance et une justesse qui l'honorent, transgresser ce rôle par des conseils judicieux et éclairés, notamment quant à la stratégie de publication à adopter.

I am very grateful to Pr. Vladimir Koltchinskii, who had the kindness to review my thesis. I am very honoured that you liked my job, and I sincerely thank you for giving me such a support.

Je remercie les autres membres de mon jury, Lucien Birgé, Olivier Catoni, Bernard Delyon, et Jian-Feng Yao de me faire l'honneur d'accepter mon invitation.

J'adresse encore de vifs remerciements à Marc Hallin, qui après avoir suivi mon premier exposé dans la communauté, aux "deuxièmes rencontres des jeunes statisticiens"en septembre 2007, est venu me féliciter et m'encourager pour la suite de ma thèse. Cet acte d'une grande générosité a été libérateur pour moi qui ne pouvais juger la valeur de mon travail, m'a donné un élan nouveau et a signé le début de mes nuits blanches !...

Je remercie plus largement les nombreuses personnes du métier avec qui j'ai pu échanger pendant cette thèse, et qui se reconnaîtront. Qu'ils me pardonnent de ne pas les nommer.

Mon activité d'enseignement a aussi joué un rôle de premier ordre pendant ma thèse, et je loue ici les bienfaits du système d'allocation couplée dont j'ai bénéficié, qui faisait de l'enseignement un de mes devoirs de doctorant. Les échanges que j'ai pu entretenir avec mes élèves m'ont beaucoup apporté, et parfois même à l'insu de ces derniers, puisqu'enseigner m'a tout simplement souvent permis de trouver une respiration bien salutaire dans mes activités de recherche. Je remercie donc l'ensemble de mes élèves, ainsi que mes encadrants, Michel Pierre, Arnaud Debussche, Grégory Vial et aussi Luc Bougé pour l'ENS ; Bachir Bekka et Florent Malrieu en ce qui concerne la répartition de mes enseignements à Rennes 1 lors de cette dernière année. Merci de m'avoir fait confiance pour ces cours/TD, et merci pour votre attitude à mon égard, toujours chaleureuse et bienveillante. Je remercie aussi Céline Vial, Jian-Feng Yao, Florent Malrieu, François Coquet, Arnaud Guyader, Benoît Cadre, Nathalie Krell, Mark Baker, Bernard Le Stum, Laurent Moret-Bailly, Nicoletta Tchou, Emmanuelle Martin, avec qui j'ai eu le plaisir de collaborer lors de ces enseignements. Travailler à votre contact m'a beaucoup appris.

J'ai une pensée particulière pour l'ensemble du personnel administratif de l'IRMAR et de l'ENS Cachan Bretagne. Je salue plus spécialement Christine Bardet, Claude Boschet, Marie-Emilie Hamel, Claudine Hélies, Véronique Le Goff, Marie-Annick Paulmier, Patrick Pérez, André Rebour, Hélène Rousseaux, Rose-Marie Tardif, Yann Toulan et Marie-Aude Verger avec qui j'ai eu le plaisir d'échanger plus avant, et toujours de manière amicale.

Que serais-je sans mes proches ?

Je remercie tous mes amis, qui m'ont aidé dans cette aventure doctorale, de près ou de loin. Je veux plus spécialement remercier trois d'entre eux, qui ont joué un rôle central. Tout d'abord Maher Kachour qui, de part son hospitalité, m'a considérablement simplifié la tâche dans mes aller-retours entre Rennes et Paris. Je remercie ensuite Katia Meziani, qui avec la gentillesse et la générosité qui la caractérise, m'a fourni un bureau officieux pour travailler à Paris - merci en passant, aux autres membres de ce bureau. Merci enfin à Benoît Patra, qui fut mon collocataire pendant les trois premières années de ma thèse. Sur un plan statistique, merci Benoît de m'avoir donné la contradiction sur mes a priori initiaux, un peu réducteurs quant aux aspects pratiques de cette science. Mon point de vue a bien changé depuis...

Merci à ma famille et à ma belle-famille, pour leur soutien constant pendant cette aventure, et en particulier pour leur aide précieuse à la réalisation du pot de thèse !

Merci enfin, ma douce Camille, pour tout ce que tu as fait pour moi durant cette thèse. Merci d'être là et de partager ma vie avec tant d'amour.

Table des Matières

Chapitre 1

Introduction

Always consider a problem under the minimal
structure in which it makes sense.

GUSTAVE CHOQUET,
cité par MICHEL TALAGRAND in *The Generic Chaining.*

Cette thèse est consacrée à l'étude théorique d'une méthode de calibration automatique des pénalités en sélection de modèles. Cette méthode, initialement formulée par Birgé et Massart [23], se base en pratique sur une heuristique, appelée *heuristique de pente*, qui se décline selon les trois points suivants. Premièrement, il existe une pénalité minimale définie comme le seuil maximal de pénalisation au-dessous duquel toute procédure de sélection de modèles par pénalisation se comporte "très mal", au sens où la perte en performance et la dimension du modèle sélectionné sont très supérieures à celles du modèle ayant la meilleure performance dans la collection de modèles considérée. Ce "meilleur modèle"est appelé l'oracle et est en pratique inconnu du statisticien. Deuxièmement, pour une pénalité supérieure à la pénalité minimale, la dimension du modèle sélectionné est "raisonnable"et l'estimateur associé satisfait une inégalité oracle, ce qui signifie que la performance de l'estimateur est comparable à celle de l'oracle. La pénalité minimale est donc aussi le seuil minimal de pénalisation au-dessus duquel la procédure de sélection de modèles se comporte raisonnablement bien. On déduit des deux premiers points l'existence d'un saut dans la dimension des modèles sélectionnés autour du seuil minimal de pénalisation. Troisièmement, si la pénalité considérée vaut deux fois la pénalité minimale, alors elle est "optimale", la performance de l'estimateur sélectionné étant dans ce cas équivalente à celle de l'oracle. En pratique, on retient de cette heuristique la règle suivante

$$\boxed{\text{“pénalité optimale”} = 2 \times \text{“pénalité minimale”}} \tag{1.1}$$

et on utilise le saut en la dimension des modèles sélectionnés, identifié en faisant varier un seuil de pénalisation préalablement choisi, pour estimer la pénalité minimale et donc la pénalité optimale via la formule (1.1). Le seuil de référence choisi par le statisticien peut être déterministe et basé sur des considérations *a priori* du problème statistique, ou estimé à partir des données, par exemple par des méthodes de rééchantillonnage comme proposé dans Arlot [7], [5].

Une question plus générale que permet de traiter l'heuristique de Birgé et Massart est celle de la *sélection de modèles optimale par pénalisation* : comment caractériser une pénalité optimale, c'est-à-dire une pénalité qui permet de sélectionner un estimateur ayant une performance équivalente à celle de l'oracle, en fonction des données du problème et comment

l'estimer en pratique ? Lorsqu'elle est valide, l'heuristique de pente permet d'identifier la pénalité optimale comme deux fois la pénalité minimale, et de l'estimer en utilisant le saut des dimensions sélectionnées autour de la pénalité minimale pour calibrer une *forme* de pénalité préalablement choisie. D'autres solutions existent au problème de sélection de modèles optimale par pénalisation, en particulier les pénalités de rééchantillonnage à poids échangeables et les pénalités V-fold proposées par Arlot [7], [5]. Toutefois, ces méthodes, qui permettent d'estimer la forme de la penalité optimale par sa version rééchantillonnée, ne sont connues en général qu'à une constante multiplicative près qu'il est nécessaire de calibrer en pratique. Les pénalités de rééchantillonnage d'Arlot et l'heuristique de pente de Birgé et Massart sont donc complémentaires, leur utilisation combinée offrant en pratique une procédure de pénalisation uniquement basée sur les données, sensée fournir dans les cas favorables une estimation effective et quasiment optimale de l'oracle.

Comme l'ont montré Arlot et Massart [10], l'heuristique de pente se formule naturellement dans un cadre général de M-estimation, et l'algorithme de calibration des pénalités qui en est issu se destine donc à un très large spectre d'applications. Bien que récente, la méthode a déjà montré son efficacité pratique dans des domaines applicatifs très variés. Ainsi, des résultats concluants ont été établis par simulation dans des contextes tels que l'estimation de réserves pétrolières (Lepez [54]), la détection de ruptures (Lebarbier [52]), la génétique (Villers [87]), les modèles de mélange (Maugis et Michel [63]), la classification non-supervisée (Baudry [20]), ou encore l'estimation de modèles graphiques (Verzelen [86]).

Cependant, la délimitation théorique du champ de validité de l'heuristique pente, qui permettrait d'éclairer dans une large mesure son efficacité pratique, demeure à l'heure actuelle un défi mathématique. En effet, l'analyse de l'heuristique de pente se base sur des contrôles très fins des quantités en jeu à modèle fixé, ce qui requiert une forte spécification des structures dans les problèmes abordés. Plus précisement, la pierre angulaire de l'heuristique de pente réside dans l'équivalence présumée de l'*excès de risque* - qui mesure la performance d'un M-estimateur - avec sa contrepartie empirique, appelée *excès de risque empirique*, pour les M-estimateurs susceptibles d'être sélectionnés parmi la collection considérée. Pour démontrer un tel fait, la stratégie généralement adoptée est d'obtenir un contrôle *à la constante près* et avec grande probabilité de l'excès de risque et de l'excès de risque empirique sur un modèle fixé, puis d'en déduire que la différence de ces deux quantités est négligeable, uniformément sur l'ensemble des modèles de dimension "raisonnable"dans la collection considérée. Ceci suggère en particulier d'établir des bornes inférieures et supérieures de déviation, optimales au premier ordre, pour l'excès de risque et pour sa contrepartie empirique. Bien que les bornes supérieures de concentration de l'excès de risque aient été largement étudiées dans des contextes généraux d'estimation non-paramétrique par minimum de contraste, et en lien avec le développement de la théorie statistique de l'apprentissage, la question des bornes inférieures de déviation pour l'excès de risque d'un M-estimateur dans un cadre général reste quasiment vierge dans la littérature. De plus, des constantes optimales dans les bornes de déviation des excès de risque n'ont été exhibées que dans des travaux récents dédiés à l'heuristique de pente ou à la sélection de modèles optimale dans des cadres bien spécifiques, tels que la régression homoscédastique avec bruit Gaussien homoscédatique (Birgé et Massart [23] puis Baraud, Giraud et Huet [11]), l'estimation de la densité par maximum de vraisemblance sur des histogrammes (Castellan [30]), la régression hétéroscédastique avec un *design* aléatoire sur des modèles par histogrammes (Arlot et Massart [10], Arlot [7], [5]) et l'estimation de la densité par moindres carrés sur des modèles linéaires (Lerasle [56], [55]).

Le point commun de toutes ces études est qu'elles se basent sur une écriture *explicite* des estimateurs en fonction des données du problème considéré. Les fonctions constantes par morceaux sur une partition donnée ont par exemple la propriété remarquable d'être engendrées par une famille d'indicatrices à supports disjoints, et donc orthogonales entre elles pour toute loi sous-jacente et en particulier pour la loi empirique des données, ce qui permet une

écriture simple de l'estimateur des moindres carrés dans cette base. En régression avec bruit hétéroscédastique et *design* aléatoire sur un modèle par histogrammes, on peut ainsi calculer de manière exacte l'espérance de l'excès de risque empirique et la stratégie adoptée dans ce cadre par Arlot et Massart [10], tirant profit de cette information, est en substance la suivante : en premier lieu, les auteurs établissent des inégalités de concentration pour l'excès de risque et pour l'excès de risque empirique. Ils donnent ensuite un encadrement fin de l'espérance de l'excès de risque en fonction de l'espérance de l'excès de risque empirique. Enfin, l'équivalence de l'excès de risque et de l'excès de risque empirique est déduite des résultats précédents en montrant que les espérances respectives de ces quantités sont équivalentes et que leurs déviations sont négligeables devant l'espérance de l'excès de risque empirique.

La concentration de l'excès de risque empirique à modèle fixé est démontrée par Arlot et Massart [10] en utilisant les résultats obtenus dans [27] par Boucheron et Massart, dans un contexte très général de M-estimation bornée avec conditions de marge. Boucheron et Massart [27] montrent en effet que la concentration de l'excès de risque est un phénomène général lié à la M-estimation, sous des conditions très souples décrivant la richesse du modèle considéré en termes d'incréments locaux maximaux moyens du processus empirique indexé par les fonctions de perte associées au modèle. Ce cadre permet notamment de traiter le cas de la classification binaire sur une classe de Vapnik-Červonenkis, et ainsi d'obtenir des inégalités de concentration pour l'excès de risque empirique qui s'expriment en fonction des hypothèses de marge. Les résultats de Boucheron et Massart [27] sont donc un outil central pour aborder l'heuristique de pente, et soutiennent la généralité de cette heuristique. Toutefois, pour obtenir un contrôle par bornes inférieures et supérieures de l'excès de risque empirique avec grande probabilité, les inégalités de concentration de Boucheron et Massart [27] doivent être associées à un encadrement de l'espérance de l'excès de risque. Dans un cadre général, un tel encadrement reste une question ouverte. Ainsi, la validation de l'heuristique de pente dans le cas de la classification binaire demeure un problème ouvert à l'heure actuelle, d'un intérêt majeur en apprentissage statistique, puisqu'elle permettrait sûrement d'apporter des réponses décisives en pratique au problème de sélection de prédicteurs adaptatifs à la marge.

Dans cette thèse, notre apport personnel à l'étude théorique de l'heuristique de pente réside dans la définition d'un cadre général, qui s'inscrit dans le contexte de la M-estimation et que nous appelons "estimation par minimum de contraste régulier", et dans la validation de l'heuristique de pente dans ce cadre, sous des hypothèses génériques sur la collection de modèles considérée. Pour ce faire, nous développons une méthodologie de preuve inédite permettant de traiter à la fois le problème des bornes supérieures de déviation pour les excès de risque et le problème des bornes inférieures de déviation pour ces mêmes quantités, et donnant des résultats optimaux au premier ordre dans le cadre des contrastes réguliers. Cette approche permet de se libérer de l'utilisation de formules explicites pour les M-estimateurs considérés, et seule leur caractérisation implicite comme minimiseurs du risque empirique est utilisée. La méthode développée ouvre aussi une voie de recherche pour des cadres de M-estimation qui ne sont pas forcément à contraste régulier, comme par exemple la classification binaire.

Nous étudions trois exemples d'estimation par minimium de contraste régulier, à savoir la régression par moindres carrés, avec bruit hétéroscédastique et *design* aléatoire sur des modèles linéaires, l'estimation de la densité par moindres carrés, sur des modèles linéaires ou affines, et enfin l'estimation de la densité par maximum de vraisemblance sur des modèles convexes lorsqu'une version du théorème de Pythagore est vérifée pour la divergence de Kullback-Leibler sur ces modèles. Ceci nous permet aux Chapitres 3 et 4 de retrouver des résultats similaires à Arlot et Massart [10], dans le cas de la régression bornée sur des modèles par histogrammes et sous le même jeu d'hypothèses que dans [10], et de les étendre en particulier au cas des polynômes par morceaux. Nos résultats montrent donc, comme conjecturé dans Arlot et Massart [10], que l'heuristique de pente est valide en régression hétéroscédastique avec *design* aléatoire, pour des modèles linéaires plus généraux que les modèles par histogrammes. Nous

montrons au Chapitre 6 que notre approche permet de retrouver des résultats similaires à ceux de Lerasle [56], pour le contrôle de la pénalité optimale. Nous validons au Chapitre 5 l'heuristique de pente dans le cas de l'estimation de la densité par maximum de vraisemblance, pour le risque de Kullback-Leibler, sur des modèles par histogrammes, affinant ainsi les résultats obtenus précédemment dans ce cadre par Castellan [30]. Ces résultats donnent, à notre connaissance, pour la première fois, la validité de l'heuristique de pente pour un risque non quadratique. Le Chapitre 2 est dédié à la notion de contraste régulier. Nous établissons au Chapitre 7 des bornes inférieures et supérieures de déviations à modèle fixé pour l'excès de risque et sa contrepartie empirique dans le cadre général de l'estimation par minimum de contraste régulier. Ces bornes sont optimales au premier ordre, et permettent en particulier un contrôle à la constante près des excès de risque, montrant ainsi leur équivalence pour des modèles de dimension raisonnable. Nous donnons enfin des perspectives de recherche dans le dernier chapitre de cette thèse.

Loin de mettre un point final à la question de la validité de l'heuristique de pente, nos méthodes se basent fortement sur la linearité des modèles considérés. En particulier, le cas de grandes collections de modèles - c'est-à-dire des collections avec un nombre au moins exponentiel de modèles -, qu'il est nécessaire d'envisager dans des contextes tels que la sélection complète de variables ou la détection de ruptures multiples, reste à l'heure actuelle et même dans le cadre de l'estimation par minimum de contraste régulier, un problème ouvert. En effet, pour pouvoir définir une pénalité optimale dans ce cas, qui permette d'obtenir l'équivalence du risque de l'estimateur sélectionné avec celui de l'oracle, il est nécessaire de "regrouper"les modèles de "complexités"équivalentes au regard du problème posé, et il s'en suit en général, la perte de la linéarité sur ces unions de modèles. Des bornes inférieures de pénalité n'ont pu être obtenues dans le cas de grandes collections de modèles que dans des contextes Gaussiens, par Birgé et Massart [23] dans un cadre qui comprend en particulier la régression avec *design* fixe et bruit homoscédastique, puis pas Baraud, Giraud et Huet [11] considérant de plus que le niveau bruit est inconnu, et fournissant des pénalités prenant en compte son estimation.

1.1 Le problème général de M-estimation

Nous donnons dans cette section une formulation générale et inédite du problème de M-estimation. Nous définissons ainsi les quantités fondamentales intervenant en M-estimation et les illustrons par quelques exemples classiques. Une introduction générale à la M-estimation est disponible dans le livre de van de Geer [77], consacré à l'étude de ce cadre statistique par des méthodes de processus empirique. On pourra aussi consulter le livre de Massart [61], dédié à la sélection de modèles pour les M-estimateurs, d'un point de vue non asymptotique se basant notamment sur des inégalités de concentration pour les processus Gaussiens et empiriques.

Soit $(\mathcal{Z}, \mathcal{T})$ un espace mesurable et μ une loi de probabilité sur $(\mathcal{Z}, \mathcal{T})$. On considère l'échantillon $\xi_1, ..., \xi_n$ de n variables aléatoires de même loi de probabilité P sur $(\mathcal{Z}, \mathcal{T})$. On définit aussi ξ, une variable aléatoire générique de loi P indépendante de $(\xi_1, ..., \xi_n)$. On note les espérances comme suit : pour une fonction f convenable,

$$Pf = P(f) = \mathbb{E}[f(\xi)]$$
$$\mu f = \mu(f) = \int_{\mathcal{Z}} f d\mu$$

et pour

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$$

la loi empirique des données $(\xi_1, ..., \xi_n)$, on note

$$P_n f = P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(\xi_i) \ .$$

La variance de $\sqrt{n} \cdot P_n f$ est

$$\mathrm{Var}(f) := \mathbb{V}[f(\xi)] = \mathbb{E}\left[f^2(\xi)\right] - (\mathbb{E}[f(\xi)])^2 \ .$$

La partie positive d'un nombre réel $x \in \mathbb{R}$ est notée $(x)_+ := \max\{x\ ,\ 0\} \geq 0$ et sa partie négative est $(x)_- := (-x)_+ = \max\{-x\ ,\ 0\} \geq 0$. Nous étendons ces définitions aux fonctions réelles f definies sur $\mathcal{Z}$ de la manière suivante,

$$(f)_+ : z \in \mathcal{Z} \longmapsto (f(z))_+ \quad , \quad (f)_- : z \in \mathcal{Z} \longmapsto (f(z))_- \ .$$

On note aussi $L_1^-(P)$ l'ensemble des fonctions réelles mesurables sur $(\mathcal{Z}, \mathcal{T})$ de partie négative intégrable pour la loi P,

$$L_1^-(P) = \left\{f : \mathcal{Z} \longrightarrow \mathbb{R} \ \ \mathcal{T}\text{-mesurable}\ ;\ P(f)_- < +\infty\right\} \ .$$

On remarque alors que l'on peut définir convenablement l'espérance des fonctions $f \in L_1^-(P)$ pour la loi P, et pour toute fonction $f \in L_1^-(P)$, on pose

$$Pf := P(f)_+ - P(f)_- \in \overline{\mathbb{R}}\ ,$$

où $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$.

Définition *Une fonctionnelle K d'un espace de fonctions $\mathcal{S}$ vers $L_1^-(P)$,*

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow \left\{f : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R}\ ;\ P(f)_- < +\infty\right\} \\ s \longmapsto (Ks : z \longmapsto (Ks)(z)) \end{array} \right. ,$$

est appelée ***contraste*** *s'il existe un unique élément $s_* \in \mathcal{S}$ tel que*

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) \quad \text{et} \quad P(Ks_*) < +\infty \ . \tag{1.2}$$

La fonction s_* est appelée la ***cible****. Pour tout $s \in \mathcal{S}$, Ks est une* ***fonction contrastée*** *et $P(Ks) \in \overline{\mathbb{R}}$ est appelé le* ***risque*** *de s.*

Comme on a $(Ks_*) \in L_1^-(P)$, on voit facilement que la condition $P(Ks_*) < +\infty$ est équivalente à

$$Ks_* \in L_1(P) := \{f : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R}\ ;\ P|f| < +\infty\} \ .$$

D'après (1.2), la cible s_* est le minimiseur du risque sur $\mathcal{S}$. Cette quantité est inconnue car elle dépend de la loi P des données, et l'un des buts principaux en M-estimation est d'estimer cette quantité via les données $(\xi_1, ..., \xi_n)$. On donne maintenant la définition d'un M-estimateur, où M signifie "minimum"- ne pas confondre avec le modèle M.

Définition *Soit $(K, \mathcal{S}, P)$ un triplet tel que $K : \mathcal{S} \longrightarrow L_1^-(P)$ soit un contraste. On choisit $M \subset \mathcal{S}$. M est alors appelé un* ***modèle****. Un* ***M-estimateur*** *$s_n(M)$, associé au modèle M pour le contraste K, est défini par*

$$s_n(M) \in \arg\min_{s \in M} P_n(Ks) \quad \text{and} \quad P_n(Ks_n(M)) < +\infty \ \ p.s., \tag{1.3}$$

où la quantité $P_n(Ks)$ *est appelée le* ***risque empirique*** *de* s*. L'existence de* $s_n(M)$ *n'étant pas garantie, la définition suivante peut se révéler pratique. Pour tout* $\rho > 0$*, l'ensemble* $\mathcal{V}_n(\rho, M)$ *des* ρ**-minimiseurs empiriques** *sur* M *est*

$$\mathcal{V}_n(\rho, M) := \left\{ s \in M \ ; \ P_n(Ks) \leq \inf_{t \in M} P_n(Kt) + \rho \right\} . \tag{1.4}$$

Notons que $P_n(Ks)$ est bien défini pour tout $s \in M$. En effet pour $s \in M$, on a aussi $s \in L_1^-(P)$ et donc $-\infty < (Ks)(\xi_i) \leq +\infty$ P-*p.s.* pour tout $i \in \{1, ..., n\}$. De plus, la condition $P_n(Ks_n(M)) < +\infty$ *p.s.* est équivalente à $Ks_n(M) \in L_1(P_n)$ et assure que l'on ne soit pas dans le cas dégénéré où pour tout $s \in M$, $P_n(Ks) = +\infty$.
D'après (1.3), un M-estimateur $s_n(M)$ vise à estimer la cible s_* en minimisant la contrepartie empirique du critère (1.2) définissant s_*, sur un certain sous-ensemble M de $\mathcal{S}$. Une des tâches principales du statisticien est alors de choisir un "bon"modèle M pour estimer s_*, et les méthodes de sélection de modèles visent à automatiser cette tâche pour une *collection* donnée de modèles. La définition suivante fournit un critère naturel de performance d'un M-estimateur, et en particulier un critère "idéal"de sélection de modèles.

Définition *Soit* $(K, \mathcal{S}, P)$ *un triplet, avec* $K : \mathcal{S} \longrightarrow L_1^-(P)$ *un contraste, de cible associée* s_**. L'****excès de risque*** $\ell(s_*, s)$ *d'une fonction* $s \in \mathcal{S}$ *est défini par*

$$\ell(s_*, s) := P(Ks) - P(Ks_*) = P(Ks - Ks_*) \geq 0 . \tag{1.5}$$

Pour un modèle $M \subset \mathcal{S}$*, si l'on suppose qu'un M-estimateur* $s_n(M)$ *existe sur* M*, alors l'excès de risque de* $s_n(M)$*, aussi appelé le* ***vrai excès de risque****, est la quantité aléatoire*

$$\ell(s_*, s_n(M)) = P(Ks_n(M) - Ks_*) \geq 0 . \tag{1.6}$$

On remarque que la quantité $\ell(s_*, s) \in \overline{\mathbb{R}}_+ := \mathbb{R}_+ \cup \{+\infty\}$ est bien définie pour tout $s \in \mathcal{S}$ puisque $Ks \in L_1^-(P)$ et $Ks_* \in L_1(P)$. Un M-estimateur est d'autant plus performant que son excès de risque est faible. Il existe une contrepartie empirique à l'excès de risque, définie comme suit.

Définition *Soit* $(K, \mathcal{S}, P)$ *un triplet, avec* $K : \mathcal{S} \longrightarrow L_1^-(P)$ *un contraste, de cible associée* s_**. Pour un modèle* $M \subset \mathcal{S}$*, si l'on suppose qu'un M-estimateur* $s_n(M)$ *existe sur* M*, alors l'****excès de risque empirique*** *de* $s_n(M)$ *est la quantité aléatoire*

$$P_n(Ks_* - Ks_n(M)) \ . \tag{1.7}$$

On remarque que l'excès de risque empirique est bien défini car $P_n(Ks_n(M)) < +\infty$, et contrairement au vrai excès de risque, il n'est pas exclu qu'il soit négatif. Par définition d'un M-estimateur, l'excès de risque empirique est une quantité croissante pour l'inclusion des modèles $M \subset \mathcal{S}$. L'excès de risque fournit donc une mesure de la "complexité"d'un modèle. Un problème central en sélection de modèles par pénalisation est de comprendre la relation qu'il peut exister entre l'excès de risque et l'excès de risque empirique.
On définit aussi les notions suivantes, associées au choix d'un modèle M.

Définition *Soit* $(K, \mathcal{S}, P)$ *un triplet, avec* $K : \mathcal{S} \longrightarrow L_1^-(P)$ *un contraste de cible associée* s_* et soit $M \subset \mathcal{S}$*. Un* ***projeté*** s_M *de la cible* s_* *sur* M*, vérifie*

$$s_M \in \arg\min_{s \in M} P(Ks) \quad \text{et} \quad P(Ks_M) < +\infty . \tag{1.8}$$

*Si un tel projeté existe, alors on définit les quantités suivantes, qui sont indépendantes du choix du projeté. L'**excès de risque sur** M pour toute fonction $s \in M$, est donné par*

$$P(Ks - Ks_M) \geq 0$$

et l'excès de risque sur M d'un M-estimateur $s_n(M) \in M$ vaut

$$P(Ks_n(M) - Ks_M) \geq 0 \ . \tag{1.9}$$

*De plus, l'**excès de risque empirique sur** M, d'un M-estimateur $s_n(M) \in M$ par rapport à un projeté s_M, est*

$$P_n(Ks_M - Ks_n(M)) \geq 0 \ .$$

Un projeté s_M de la cible s_* sur le modèle M, est donc un minimiseur du risque sur le modèle M. L'image par le contraste d'un projeté s_M a, de plus, une espérance finie sous la loi P, *i.e.* $P(Ks_M) < +\infty$. On note que comme $s_M \in M$, on a déjà $P(Ks_M)_- < +\infty$, et donc $P(Ks_M) > -\infty$. La propriété $P(Ks_M) < +\infty$ est donc équivalente à $Ks_M \in L_1(P)$ et assure que l'on ne soit pas dans le cas dégénéré où pour tout $s \in M$, $P(Ks_M) = +\infty$. Si un tel projeté s_M existe, on remarque que l'excès de risque d'un M-estimateur $s_n(M)$ se décompose en la somme de l'excès de risque du projeté s_M et de l'excès de risque sur M du M-estimateur :

$$P(Ks_n(M) - Ks_*) = \underbrace{P(Ks_n(M) - Ks_M)}_{\text{terme de "variance"}} + \underbrace{P(Ks_M - Ks_*)}_{\text{biais du modèle}} . \tag{1.10}$$

L'excès de risque d'un projeté s_M, donné par $P(Ks_M - Ks_*) = \ell(s_*, s_M)$, mesure la qualité d'approximation du modèle M pour la cible s_*, en terme de risque. Cette quantité est généralement appelée le **biais** du modèle M par rapport s_*. Comme nous le verrons aux Sections 1.3 et 1.4, l'excès de risque sur M mesure la complexité du modèle M, au regard du contraste K et de la loi P, et il est aussi traditionnellement appelé **terme de variance**. Le choix d'un "grand" modèle, s'il n'est pas abscons, mène généralement à un petit biais et un terme de variance élevé. Au contraire, un "petit"modèle aura en général un biais élevé et une petite variance. Un des buts premiers de la sélection de modèles est d'opérer, par une méthode systématique, un *compromis biais-variance* dans l'espoir de sélectionner un estimateur avec un faible risque (cf. Massart [61]).

Nous donnons maintenant quelques exemples de M-estimation, afin d'illustrer les définitions précédentes. Nous commençons par l'exemple le plus classique, dont la M-estimation est une généralisation, à savoir l'estimateur du maximum de vraisemblance.

- **Estimation de la densité par maximum de vraisemblance** : on suppose que P a une densité

  $$s_* = \frac{dP}{d\mu} \ ,$$

  pour la loi μ sur $(\mathcal{Z}, \mathcal{T})$ telle que $P(\ln s_*)_+ < +\infty$. Alors, en prenant

  $$\mathcal{S} = \left\{ s \geq 0 \ \mathcal{T}-\text{mesurable} \ ; \ \int_{\mathcal{Z}} s d\mu = 1 \text{ et } P(\ln s)_+ < +\infty \right\} \ ,$$

  avec la convention $\ln 0 = -\infty$, en définissant le contraste de Kullback-Leibler

  $$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1^-(P) \\ s \longmapsto (Ks : z \in \mathcal{Z} \mapsto -\ln(s(z))) \end{array} \right. ,$$

  il vient, par l'inégalité de Jensen,

  $$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) \ .$$

De plus, on a toujours $P(\ln s)_- < +\infty$, et donc ici $Ks_* \in L_1(P)$. Si l'on choisit un modèle $M \subset \mathcal{S}$, alors le M-estimateur $s_n(M)$, s'il existe, est l'estimateur classique du maximum de vraisemblance sur M. Dans ce contexte, pour tout $s \in \mathcal{S}$, l'excès de risque

$$\ell(s_*, s) = P(Ks - Ks_*) = \mathcal{K}(s_*, s) := \int_{\mathcal{Z}} s_* \ln\left(\frac{s_*}{s}\right) d\mu$$

n'est autre que la divergence de Kullback-Leibler de la densité s par rapport à la cible s_*. Pour un modèle $M \subset \mathcal{S}$, on peut parfois garantir l'existence et l'unicité d'un projeté s_M sur M de la cible s_*, alors appelé le projeté de Kullback, comme par exemple dans le cas des familles exponentielles. On pourra consulter l'article de Csiszár and Matúš [35], pour un point de vue récent et généralisant sur cette problématique.

- **Régression par moindres carrés** : on suppose que $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ pour un espace mesurable $\mathcal{X}$ et que pour $\xi = (X, Y) \in \mathcal{X} \times \mathbb{R}$ de loi P on a

$$Y = s_*(X) + \sigma(X)\varepsilon ,$$

avec $\mathbb{E}\left[Y^2\right] < +\infty$, $\mathbb{E}\left[\varepsilon \,|X\right] = 0$ et $\mathbb{E}\left[\varepsilon^2 \,|X\right] = 1$. Alors $s_* = \mathbb{E}\left[Y \,|X = \cdot\right]$ est la fonction de régression de Y par X. On note alors

$$\mathcal{S} = L_2\left(P^X\right) := \left\{s : \mathcal{X} \to \mathbb{R} \ ; \ \mathbb{E}\left[s^2(X)\right] < +\infty\right\}$$

et on définit le contraste des moindres carrés en régression comme étant

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1(P)\left(\subset L_1^-(P)\right) \\ s \longmapsto \left(Ks : z = (x, y) \in \mathcal{Z} \mapsto (y - s(x))^2\right) \end{array} \right. .$$

Il vient

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) \quad , \quad \ell(s_*, s) = \|s - s_*\|^2_{L_2(P^X)} := \int_{\mathcal{X}} (s - s_*)^2(x)\, dP^X$$

et l'excès de risque n'est autre que la perte quadratique dans $L_2\left(P^X\right)$. Les M-estimateurs associés au contraste des moindres carrés en régression sont les estimateurs classiques des moindres carrés. Pour un modèle linéaire, le projeté s_M n'est autre que le projeté orthogonal dans $L_2\left(P^X\right)$ de la cible dans le modèle M considéré.

L'exemple suivant sort typiquement du cadre de l'estimation par minimum de contraste régulier que nous définissons à la Section 1.2.

- **Classification binaire** : on suppose que $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ pour un espace mesurable $\mathcal{X}$ et on définit $\xi = (X, Y) \in \mathcal{X} \times \{0, 1\}$ une variable de loi P. Si on pose

$$\mathcal{S} = \{s : \mathcal{X} \longrightarrow \{0, 1\} \text{ mesurable}\} ,$$

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1(P)\left(\subset L_1^-(P)\right) \\ s \longmapsto \left(Ks : z = (x, y) \in \mathcal{Z} \mapsto \mathbf{1}_{\{y \neq s(x)\}}\right) \end{array} \right. ,$$

et

$$s_* : x \in \mathcal{X} \longmapsto \mathbf{1}_{\{\mathbb{E}[Y|X=x] \geq 1/2\}} ,$$

alors le risque $P(Ks) = \mathbb{P}(Y \neq s(X))$ est la probabilité que le "classifieur" $s \in \mathcal{S}$ prédise le mauvais label pour la variable X, et de plus on a

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) .$$

La cible s_* est appelée le classifieur de Bayes.

1.2 Apport à la M-estimation, la notion de contraste régulier

Nous avons introduit à la Section 1.1 le problème de M-estimation dans son contexte général. Nous formulons à présent des contraintes structurelles sur le contraste, définissant ainsi le cadre de l'estimation par minimum de contraste régulier. De plus, nous donnons trois exemples de contextes réguliers, dans le cas de la régression par moindres carrés, de l'estimation de la densité par moindres carrés et enfin dans le cas de l'estimation de la densité par maximum de vraisemblance. Ces exemples sont étudiés en détails aux Chapitres 3 et 4 pour la régression par moindres carrés, le Chapitre 5 est dédié quant à lui à l'estimation de la densité par maximum de vraisemblance et enfin nous examinons le cas de l'estimation de la densité par moindres carrés au Chapitre 6.

La notion de contraste régulier nous permet au Chapitre 7 d'obtenir des bornes supérieures et inférieures pour l'excès de risque et pour l'excès de risque empirique, à la constante près, en considérant des modèles linéaires ou affines.

1.2.1 Définition d'un contraste régulier

Soit

$$L_\infty(P) := \{s : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R} \;\; \mathcal{T} - \text{mesurable} \; ; \;\; \|s\|_\infty := \text{essup}_{z\in\mathcal{Z}} (|s(z)|) < +\infty\} \;\; ,$$

où le supremum essentiel essup est pris relativement à la loi P, et soit

$$L_2(P) := \left\{s : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R} \;\; \mathcal{T} - \text{mesurable} \; ; \;\; \|s\|_2 := P\left(s^2\right) < +\infty\right\} \;\; .$$

Pour un sous-ensemble $A \subseteq \mathbb{R}$, on note $\mathring{A}$ son intérieur. Un contraste régulier se définit comme suit.

Définition *Soit $(K, \mathcal{S}, P)$ un triplet, avec $K : \mathcal{S} \longrightarrow L_1^-(P)$ un contraste de cible associée s_*. Soit $M \subset \mathcal{S} \bigcap L_\infty(P)$ un modèle. Le contraste K est dit **régulier** pour le modèle M et sous la loi P si les propriétés suivantes sont vérifiées. Il existe un unique projeté s_M de s_* sur M, satisfaisant*

$$s_M = \arg\min_{s\in M} P(Ks) \quad \text{et} \quad P(Ks_M) < +\infty \;\; . \tag{1.11}$$

Pour tout $s \in M$ et pour P-presque tout $z \in \mathcal{Z}$, on a le développement suivant,

$$Ks(z) - Ks_M(z) = \psi_0^s + \psi_{1,M}(z)(s - s_M)(z) + \psi_2(\psi_{3,M}(z)(s - s_M)(z)) \tag{1.12}$$

où ψ_0^s est une constante qui dépend de s mais pas de z, $\psi_{1,M}$ et $\psi_{3,M}$ sont des fonctions définies sur $\mathcal{Z}$, indépendantes de la fonction s considérée et non identiquement nulles. De plus $\psi_{1,M} \in L_2(P)$, $\psi_{3,M} \in L_\infty(P)$ et ψ_2 est une fonction dépendante de s, définie sur un ensemble $D_2 \subseteq R$ tel que $0 \in \mathring{D}_2$, $\psi_2(\mathcal{D}_2) \subseteq \overline{\mathbb{R}}$ et $\psi_2(0) = 0$. Il existe aussi des constantes $A_2, L_2 > 0$ telles que pour tout $\delta \in [0, A_2]$, on a $[-\delta, \delta] \subset D_2$ et pour tout $(x, y) \in [-\delta, \delta]$,

$$|\psi_2(x) - \psi_2(y)| \leq L_2\delta|x - y| \;\; . \tag{1.13}$$

Enfin, en définissant

$$M_0 = Vect\{s - s_M \; ; \; s \in M\} \;\; , \tag{1.14}$$

il existe une norme Hilbertienne $\|\cdot\|_{H,M}$ sur M_0 et des constantes $A_H, L_H > 0$ telles que, pour tout $t \in M_0$,

$$\|t\|_2 \leq A_H \|t\|_{H,M} \tag{1.15}$$

et pour tout $\delta \in \left[0, L_H^{-1}\right]$ et tout $s \in M$ tels que $\|s - s_M\|_\infty \leq \delta \leq L_H^{-1}$, on a

$$(1 - L_H\delta)\|s - s_M\|_{H,M}^2 \leq P(Ks - Ks_M) \leq (1 + L_H\delta)\|s - s_M\|_{H,M}^2 \;\; . \tag{1.16}$$

Si l'on peut prendre $\psi_2 \equiv 0$ pour tout $s \in M$, alors le contraste est dit linéaire et l'inégalité (2.24) est satisfaite pour tout $A_2, L_2 > 0$, avec $D_2 = R$.

Commentons la définition précédente. Pour qu'un contraste soit régulier pour un modèle M et sous une loi P, on demande que trois propriétés soient satisfaites.

Premièrement, on demande qu'il existe un unique projeté de la cible dans le modèle considéré. Nous rediscuterons de l'unicité du projeté un peu plus bas.

Deuxièmement, on demande que le contraste K, convenablement recentré par le projeté contrasté Ks_M, puisse être développé en la somme d'un terme constant, d'une partie linéaire et d'une partie quadratique, pour toute fonction $s \in M$. La condition (1.13) et le fait que $\psi_{3,M}$ soit uniformément borné sur $\mathcal{Z}$, assurent que le terme dépendant de ψ_2 et $\psi_{3,M}$ dans le développement du contraste se comporte quadratiquement.

Troisièmement, on demande que l'excès de risque sur M soit encadré par une norme Hilbertienne $\|\cdot\|_{H,M}$ dès que les fonctions $s \in M$ considérées sont assez proches, en norme infinie, du projeté s_M. Plus précisement, l'excès de risque $P(Ks - Ks_M)$ sur M d'une fonction $s \in M$, est équivalent à la norme $\|s - s_M\|_{H,M}$ lorsque s tend vers s_M en norme infinie. De plus, la norme quadratique dans $L_2(P)$ est dominée par la norme Hilbertienne $\|\cdot\|_{H,M}$ sur l'espace vectoriel M_0, engendré par les fonctions de M recentrées par le projeté s_M. En particulier, ceci assure l'unicité "locale"du projeté s_M. En effet, soit $s \in M$ tel que $\|s - s_M\|_\infty < L_H^{-1}$, alors on a

$$P(Ks - Ks_M) \geq (1 - L_H\delta)\|s - s_M\|_{H,M}^2 \geq A_H^{-2}(1 - L_H\delta)\|s - s_M\|_2^2 > 0 ,$$

donc $P(Ks) > P(Ks_M) = \inf_{s\in M} P(Ks)$ et le projeté s_M est donc forcément unique sur

$$M \bigcap \left\{ s \in L_\infty(P) \ ; \ \|s - s_M\|_\infty < L_H^{-1} \right\} .$$

Nous décrivons à présent les trois exemples de contrastes réguliers étudiés dans cette thèse.

1.2.2 Trois exemples de contrastes réguliers

Estimation de la densité par maximum de vraisemblance sur des modèles par histogrammes

Rappelons que dans le cas de l'estimation de la densité par maximum de vraisemblance on a

$$s_* = \frac{dP}{d\mu} \ , \mathcal{S} = \left\{ s \geq 0 \ \mathcal{T}-\text{mesurable} \ ; \ \int_{\mathcal{Z}} s d\mu = 1 \text{ et } P(\ln s)_+ < +\infty \right\} ,$$

et K est ici le contraste de Kullback-Leibler

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1^-(P) \\ s \longmapsto (Ks : z \in \mathcal{Z} \mapsto -\ln(s(z))) \end{array} \right. .$$

On demande aussi $(Ks_*) \in L_1(P)$. Soit M un modèle de densités constantes par morceaux sur une partition Λ_M de $\mathcal{Z}$,

$$M = \left\{ s = \sum_{I\in\Lambda_M} \beta_I \mathbf{1}_I \ ; \ (\beta_I)_{I\in\Lambda_M} \in \mathbb{R}^{D_M}, \ s \geq 0, \ \int_{\mathcal{Z}} s d\mu = 1 \right\}$$

avec $D_M = \text{Card}(\Lambda_M)$ et pour tout $I \in \Lambda_M$, $\mu(I) > 0$. Au Chapitre 5, où ce cas est étudié en détails, M est noté $\widetilde{M}$. On vérifie aisément que le projeté s_M existe et est unique. Il est donné par

$$s_M = \sum_{I\in\Lambda_M} \frac{P(I)}{\mu(I)} \mathbf{1}_I \ . \tag{1.17}$$

On voit aisément que $Ks_M \in L_1(P)$. De plus, l'estimateur du maximum de vraisemeblance $s_n(M)$ existe et est unique, et s'écrit

$$s_n(M) = \sum_{I \in \Lambda_M} \frac{P_n(I)}{\mu(I)} \mathbf{1}_I \ .$$

On note que s'il existe $I \in \Lambda_M$, tel que $P_n(I) = 0$ et $P(I) > 0$ alors on a $P(Ks_n(M)) = +\infty$, en d'autres termes $(Ks_n(M)) \in L_1^-(P) \setminus L_1(P)$.
On introduit $\psi_{1,M}$ et $\psi_{3,M}$ deux fonctions sur $\mathcal{Z}$ satisfaisant

$$\psi_{1,M} = -\psi_{3,M} = -\frac{1}{s_M}$$

et on définit aussi

$$\psi_2 : x \in [-1; +\infty) \, (:= \mathcal{D}_2) \longrightarrow \begin{cases} x - \log(1+x) & \text{if} \quad x > -1 \\ +\infty & \text{if} \quad x = -1 \end{cases} .$$

On remarque que $0 \in \mathring{\mathcal{D}}_2$, $\psi_2(\mathcal{D}_2) \subseteq \overline{\mathbb{R}}$, $\psi_2(0) = 0$, et si on pose $A_2 = 1/2$ alors pour tout $\delta \in [0, A_2]$, on a $[-\delta, \delta] \subset \mathcal{D}_2$ et pour tout $(x, y) \in [-\delta, \delta]$,

$$|\psi_2(x) - \psi_2(y)| \leq L_2 \delta |x - y| \ ,$$

avec $L_2 = 1$. De plus, $(\ln s_M) \in L_1(P)$ donc on a $s_M > 0$ P-$p.s.$, et pour tout $s \in M$, on écrit, avec la convention $\ln(0) = -\infty$,

$$\begin{aligned} Ks(z) - Ks_M(z) &= -\ln\left(\frac{s(z)}{s_M(z)}\right) = -\ln\left(1 + \frac{s(z) - s_M(z)}{s_M(z)}\right) \\ &= -\frac{s(z) - s_M(z)}{s_M(z)} + \left(\frac{s(z) - s_M(z)}{s_M(z)} - \ln\left(1 + \frac{s(z) - s_M(z)}{s_M(z)}\right)\right) \\ &= \psi_{1,M}(z)(s - s_M)(z) + \psi_2(\psi_{3,M}(z)(s - s_M)(z)) \quad P\text{-}p.s. \end{aligned}$$

Ainsi, le développement donné en (1.12) est vérifié, avec $\psi_0^s = 0$ pour tout $s \in M$. De plus, on dispose d'une propriété d'orthogonalité pour la divergence de Kullback-Leibler sur M (cf. Proposition 5.1, Chapitre 5). Plus précisement,

$$\mathcal{K}(s_*, s) = \mathcal{K}(s_*, s_M) + \mathcal{K}(s_M, s) \ \text{, pour tout } s \in M \ . \tag{1.18}$$

On montre ainsi (cf. Lemma 5.4, Chapitre 5) que s'il existe $A_{\min} > 0$ telle que $\inf_{z \in \mathcal{Z}} s_*(z) \geq A_{\min} > 0$ alors par (1.17) on a $\inf_{z \in \mathcal{Z}} s_M(z) \geq A_{\min} > 0$ et si on pose $L_H = \frac{4}{3A_{\min}} > 0$, pour tout $s \in M$ tel que $\|s - s_M\|_\infty \leq \delta \leq L_H^{-1}$, il vient

$$(1 - L_H\delta)\frac{1}{2}\left\|\frac{s - s_M}{s_M}\right\|_2^2 \leq P(Ks - Ks_M) \leq (1 + L_H\delta)\frac{1}{2}\left\|\frac{s - s_M}{s_M}\right\|_2^2 \ .$$

Ainsi, en posant

$$\|s\|_{H,M} = \frac{1}{\sqrt{2}}\left\|\frac{s}{s_M}\right\|_2 \quad \text{pour tout } s \in L_2(P) \ ,$$

il s'ensuit, comme $\inf_{z \in \mathcal{Z}} s_M(z) \geq A_{\min} > 0$, que $\|\cdot\|_{H,M}$ est une norme Hilbertienne sur $L_2(P)$ et en particulier sur M_0. Cette norme est généralement appelée la norme du Khi-deux. Enfin, si $\|s_*\|_\infty < +\infty$ on a par (1.17), $\|s_M\|_\infty \leq \|s_*\|_\infty < +\infty$, et donc pour tout $s \in L_2(P)$,

$$\|s\|_2 \leq A_H \|s\|_{H,M} \ ,$$

avec $A_H = \|s_*\|_\infty$.
Des calculs précédents, on conclut que si

$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_M(z) \leq \|s_*\|_\infty < +\infty ,$$

alors le contraste de Kullback-Leibler est régulier sur le modèle de densités constantes par morceaux M sous la loi P. Pour étendre ce résultat au contexte plus général des modèles *convexes*, nous devons en particulier généraliser la relation d'orthogonalité donnée en 1.18 (cf. Section 5.4.1, Chapitre 5 pour plus de détails sur cette question).

Régression par moindres carrés

En régression par moindres carrés on a $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ et

$$Y = s_*(X) + \sigma(X)\varepsilon ,$$

avec $\mathbb{E}\left[Y^2\right] < +\infty$, $\mathbb{E}\left[\varepsilon \,|X\right] = 0$ et $\mathbb{E}\left[\varepsilon^2 \,|X\right] = 1$. De plus, $\mathcal{S} = L_2\left(P^X\right)$, et le contraste des moindres carrés en régression est

$$K : \begin{cases} \mathcal{S} \longrightarrow L_1(P)\left(\subset L_1^-(P)\right) \\ s \longmapsto \left(Ks : z = (x,y) \in \mathcal{Z} \mapsto (y - s(x))^2\right) \end{cases} .$$

L'excès de risque est alors donné par la norme Hilbertienne dans $L_2\left(P^X\right)$,

$$\ell(s_*, s) = \|s - s_*\|^2_{L_2(P^X)} \quad \text{pour tout } s \in \mathcal{S}.$$

Par abus de notation, on identifie s définie de $\mathcal{X}$ dans $\mathbb{R}$ avec son prolongement $\tilde{s}$ à $\mathcal{Z}$, défini par

$$\tilde{s} : z = (x,y) \in \mathcal{Z} \longrightarrow \tilde{s}(z) = s(x) .$$

On peut ainsi écrire, pour tout $s \in \mathcal{S}$,

$$\ell(s_*, s) = \|s - s_*\|_2^2 .$$

On considère maintenant un sous-espace linéaire M de $L_2\left(P^X\right)$ de dimension finie. Il existe alors un unique projeté orthogonal s_M de s_* sur M, et par le théorème de Pythagore on a, pour tout $s \in M$,

$$\|s - s_*\|_2^2 = \|s - s_M\|_2^2 + \|s_M - s_*\|_2^2 . \tag{1.19}$$

On déduit donc de (1.19) que s_M est le projeté de s_* dans M au sens du risque, car d'après (1.19) il vient,

$$\begin{aligned} s_M &= \arg\min_{s \in M} \|s - s_*\|_2^2 \\ &= \arg\min_{s \in M} P(Ks - Ks_*) \\ &= \arg\min_{s \in M} P(Ks) . \end{aligned}$$

On définit alors pour tout $z = (x,y) \in \mathcal{Z}$,

$$\psi_{1,M}(z) = -2(y - s_M(x)) \ , \ \psi_{3,M}(z) = 1$$

et

$$\text{pour tout } u \in \mathbb{R} =: \mathcal{D}_2, \ \psi_2(u) = u^2 . \tag{1.20}$$

Ainsi, d'après (1.20), on a

$$0 \in \mathring{\mathcal{D}}_2 \ , \ \psi_2 (\mathcal{D}_2) \subseteq \mathbb{R} \ , \ \psi_2 (0) = 0 \ ,$$

et pour tout $A_2 > 0$, en posant $L_2 = 2A_2$, on obtient, pour tout $\delta \in [0, A_2]$ et pour tout $(x, y) \in [-\delta, \delta]$,

$$|\psi_2 (x) - \psi_2 (y)| \leq L_2 \delta |x - y| \ .$$

En outre, le contraste se développe, pour tout $s \in M$ et tout $z = (x, y) \in \mathcal{Z}$, de la manière suivante,

$$\begin{aligned} Ks (z) - Ks_M (z) &= (y - s (x))^2 - (y - s_M (x))^2 \\ &= \psi_{1,M} (z) (s - s_M) (z) + \psi_2 (\psi_{3,M} (z) (s - s_M) (z)) \ . \end{aligned}$$

On retrouve donc bien le développement donné en (1.12), avec $\psi_0^s = 0$ pour tout $s \in M$. Par le théorème de Pythagore on a également

$$P (Ks - Ks_M) = \|s - s_M\|_2^2 \ ,$$

et en posant $\|\cdot\|_{H,M} = \|\cdot\|_2$, l'inégalité (1.16) est satisfaite pour tout $L_H > 0$ et $A_H \geq 1$. Finalement, on conclut que le contraste des moindres carrés en régression est régulier pour le modèle M et sous la loi P. Cet exemple est étudié en détails aux Chapitres 3 et 4.

Estimation de la densité par moindres carrés

Soit μ une loi de probabilité connue sur $(\mathcal{Z}, \mathcal{T})$, on suppose que P admet une densité f par rapport à μ :

$$f = \frac{dP}{d\mu} \ .$$

On définit $L_2 (\mu)$, l'espace des fonctions de carré intégrable par rapport à μ, à savoir

$$L_2 (\mu) = \left\{ s \ ; \ \mu \left(s^2 \right) < +\infty \right\} \ ,$$

que l'on munit de son produit scalaire usuel

$$\langle s, t \rangle = \mu (s \cdot t) = \int_{\mathcal{Z}} s \cdot t d\mu$$

et on note $\|.\|$ la norme Hilbertienne associée, définie par

$$\|s\|^2 = \|s\|_{L_2(\mu)}^2 = \langle s, s \rangle = \mu \left(s^2 \right) = \int_{\mathcal{Z}} s^2 d\mu \ .$$

On suppose de plus, qu'il existe une fonction s_0, typiquement $s_0 \equiv 1$ si $\mathcal{Z}$ est l'intevalle $[0, 1]$, ou $s_0 \equiv 0$, on définit la cible s_* par

$$f = s_0 + s_* \ \text{ avec } \int_{\mathcal{Z}} s_* \cdot s_0 d\mu = 0 \ .$$

On définit également l'espace orthogonal à s_0 dans $L_2 (\mu)$,

$$\{s_0\}^\perp := \{s \in L_2 (\mu) \ ; \ \langle s, s_0 \rangle = 0\} \ .$$

On a donc $s_* \in \{s_0\}^\perp$. Soit $s \in \{s_0\}^\perp$, on a

$$\begin{aligned} \|s - s_*\|^2 &= \|s\|^2 - 2 \langle s, s_* \rangle + \|s_*\|^2 \\ &= \|s\|^2 - 2 \langle s, f \rangle + \|s_*\|^2 \\ &= \|s\|^2 - 2Ps + \|s_*\|^2 \end{aligned}$$

et on déduit

$$s_* = \arg\min_{s\in\mathcal{S}} P(Ks)$$

où $\mathcal{S} := \{s_0\}^\perp$ et $K : L_2(\mu) \longrightarrow L_1(P)$ est le contraste des moindres carrés en densité, vérifiant

$$Ks = \|s\|^2 - 2s \ , \text{ pour tout } s \in L_2(\mu) .$$

Soit $M \subset \mathcal{S}$ un modèle linéaire de dimension finie. Pour tout $s \in M$,

$$\langle s, s_0\rangle = \int_{\mathcal{Z}} s \cdot s_0 d\mu = 0 \ .$$

L'estimateur considéré sur M est l'estimateur des moindres carrés, défini par

$$\begin{aligned} s_n &\in \arg\min_{s\in M} P_n(Ks) \\ &= \arg\min_{s\in M} \left\{ \|s\|^2 - 2P_n s \right\} \ . \end{aligned}$$

On vérifie aisément que l'estimateur des moindres carrés existe et est unique dans ce cas. Plus précisément, si D est la dimension linéaire de $(M, \|\cdot\|)$, alors pour toute famille $(\varphi_k)_{k=1}^D$, base orthonormée de $(M, \|\cdot\|)$, l'estimateur s'écrit

$$s_n = \sum_{k=1}^{D} P_n(\varphi_k)\, \varphi_k \ .$$

Pour tout $s \in \{s_0\}^\perp$, on a

$$\begin{aligned} P(Ks - Ks_*) &= PKs - PKs_* \\ &= \|s\|^2 - 2\langle s, f\rangle - \|s_*\|^2 + 2\langle s_*, f\rangle \\ &= \|s\|^2 - 2\langle s, s_*\rangle + \|s_*\|^2 \\ &= \|s - s_*\|^2 \geq 0 \ , \end{aligned}$$

et donc l'excès de risque $P(Ks - Ks_*)$ n'est autre que la perte quadratique dans $L_2(\mu)$. Si on note s_M le projeté orthogonal de s_* sur M dans $L_2(\mu)$, on a alors

$$P(Ks_M) - P(Ks_*) = \inf_{s\in M} \{P(Ks) - P(Ks_*)\} \ , \tag{1.21}$$

et on déduit de (1.21) que s_M est l'unique projeté de s_* sur M au sens du risque,

$$s_M = \arg\min_{s\in M} P(Ks) \ .$$

Par le théorème de Pythagore on a de plus, pour tout $s \in M$,

$$\|s - s_*\|^2 = \|s - s_M\|^2 + \|s_M - s_*\|^2 \ ,$$

et il vient

$$P(Ks - Ks_M) = \|s - s_M\|^2 \geq 0 \ ,$$

pour tout $s \in M$. On pose donc $\|\cdot\|_{H,M} = \|\cdot\|$, et on vérifie aisément que l'encadrement (1.16) est vérifié pour tout $L_H > 0$. Si on suppose maintenant que $\|f\|_\infty < +\infty$, alors on a pour tout $s \in M$,

$$\|s\|_2 \leq A_H \|s\|_{H,M}$$

avec $A_H = \|f\|_\infty$. Finalement, en posant

$$\psi_{1,M} \equiv -2$$
$$\psi_0^s = \|s\|^2 - \|s_M\|^2$$

on écrit, pour tout $s \in M$, et tout $z \in \mathcal{Z}$,

$$Ks\,(z) - Ks_M\,(z) = \psi_0^s + \psi_{1,M}\,(z)\,(s - s_M)\,(z) \ .$$

Des calculs précédents, on déduit que si $\|f\|_\infty < +\infty$, alors le contraste des moindres carrés en densité est linéaire pour le modèle M sous la loi P. Ce cadre est étudié au Chapitre 6, où l'on envisage aussi le cas où f est seulement supposée appartenir à $L_2\,(\mu)$.

1.3 Bornes supérieures pour l'excès de risque en M-estimation, à modèle fixé

Pour un modèle fixé M, le terme de biais étant déterministe, l'étude des fluctuations aléatoires de l'excès de risque d'un M-estimateur se ramène au contrôle de l'excès de risque sur le modèle M. Remarquons alors que, par définition d'un M-estimateur, on a $P_n\,(Ks_n\,(M) - Ks_M) \leq 0$ et donc

$$P\,(Ks_n\,(M) - Ks_M) \leq (P - P_n)\,(Ks_n\,(M) - Ks_M) \tag{1.22}$$
$$\leq \sup_{s \in M} |(P - P_n)\,(Ks - Ks_M)| \ . \tag{1.23}$$

L'excès de risque sur le modèle M du M-estimateur est donc contrôlé par l'écart uniforme entre la loi inconnue P et la mesure empirique P_n, sur un ensemble - généralement infini - de fonctions. L'ensemble de fonctions considéré ici est

$$\{Ks - Ks_M \ ; \ s \in M\} \ .$$

Pour une classe de fonctions $f \in \mathcal{F}$, notons

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |(P_n - P)\,(f)| \tag{1.24}$$

le supremum du processus empirique sur la classe $\mathcal{F}$.
L'exploitation du contrôle général décrit en (1.23) a commencé dans le cadre de la reconnaissance de forme dans l'article fondateur de Vapnik et Červonenkis [83] (pour plus de références, voir par exemple l'ouvrage de Vapnik [82]) et l'analyse de la quantité définie en (1.24) a débuté par l'étude du problème de Glivenko-Cantelli, sur la convergence vers 0 de $\|P_n - P\|_{\mathcal{F}}$. Cette étude a permis à Vapnik et Červonenkis de dégager des caractéristiques importantes de complexité d'une classe de fonctions (ou d'ensembles), comme la VC-dimension ou l'entropie aléatoire, qui permettent de contrôler le supremum du processus empirique. À ces découvertes s'ajoutèrent dans les années 60 et 70 le développement des théorèmes limites classiques dans les espaces de Banach, qui débouchèrent sur la théorie générale des processus empiriques et l'article pionnier de Dudley [36] sur les théorèmes limites centraux pour les mesure empiriques (voir Dudley [37], Pollard [64] et van der Waart et Wellner [81]). Ces derniers résultats sont des principes d'invariance, dits faibles car en loi, puisqu'ils ramènent l'étude en loi du processus empirique à l'infini à celle d'un processus Gaussien de même covariance sur des classes générales de fonctions appelées classes de Donsker. Toutefois, ces résultats sont asymptotiques et ne permettent donc pas une analyse non-asymptotique en M-estimation.

Un outil fondamental pour l'approche non-asymptotique en statistique est la théorie de la concentration de la mesure sur les espaces produits développée par Talagrand dans les années 90,

et plus précisément une version uniforme de l'inégalité de Bernstein, qui décrit la concentration de $\|P_n - P\|_{\mathcal{F}}$ autour de sa moyenne, démontrée à l'origine par des arguments d'isopérimétrie (voir Talagrand [73] et [74]). Ces inégalités ont ensuite été redémontrées par Ledoux [72] par la méthode d'entropie (voir Ledoux [72]). Cette méthode a par la suite permis à Bousquet [28] d'obtenir les constantes optimales pour la concentration à droite du processus empirique, et Klein [42] puis Klein et Rio [41] ont obtenu des constantes quasiment optimales pour la concentration à gauche.

En M-estimation générale, ces outils ont permis d'obtenir des vitesses de convergence, appelées vitesses rapides, plus fines dans beaucoup de cas que les bornes initiales de Vapnik et Červonenkis, et souvent minimax pour les problèmes considérés, parfois à un facteur logarithmique près. Pour un rappel sur la théorie minimax, le lecteur pourra se référer par exemple à Tsybakov [76]. On peut trouver des exemples de telles bornes dans Massart [61], et plus précisement dans le chapitre 8 consacré à l'apprentissage statistique, reprenant en particulier certains résultats de Massart et Nédélec [62]. Dans ce dernier article, on trouvera une analyse minimax des bornes générales de convergence, avec un exemple approfondi montrant qu'il n'est pas possible en général d'enlever le facteur logarithmique dans les vitesses obtenues dans [62] sous des conditions générales d'entropie à crochet et aussi pour des classes de dimension de Vapnik et Červonenkis finie. La démonstration du théorème principal repose entre autre sur le contrôle global d'un processus empirique renormalisé. Par une méthode légèrement différente, qui consiste à contrôler le processus empirique renormalisé sur des couches localisées en excès de risque de la classe des fonctions considérées et à tirer partie de cette découpe par l'utilisation de nombres d'entropie prenant en compte la taille de l'enveloppe des fonctions pour la norme $L_2(P)$, Koltchinskii et Giné [39] ont pour leur part réussi à enlever le facteur logarithmique dans les cas où l'enveloppe d'une couche d'excès de risque se localise elle aussi convenablement. Ces analyses fines se basent sur des inégalités de concentration de type Talagrand et des méthodes de chaînage. Elles reposent en fait sur une propriété plus précise que l'inégalité (1.23). En effet, définissons la quantité $\phi_n(\delta)$ par

$$\sup_{\{s\in M; P(Ks-Ks_M)\leq\delta\}} |(P-P_n)(Ks-Ks_M)| \leq \phi_n(\delta) \ . \tag{1.25}$$

L'ensemble $\{s \in M : P(Ks - Ks_M) \leq \delta\}$ est parfois appelé l'ensemble δ-minimal, c'est en effet l'ensemble des fonctions du modèle M dont le risque est à une distance inférieure à δ du risque minimal sur M. On a alors par (1.22),

$$P(Ks_n(M) - Ks_M) \leq \phi_n(P(Ks_n(M) - Ks_M)) \ . \tag{1.26}$$

Une borne supérieure pour l'excès de risque est alors fournie par un majorant de la plus grande solution de l'inéquation

$$\delta \leq \phi_n(\delta) \ . \tag{1.27}$$

Cependant, pour obtenir une vitesse déterministe, le contrôle en (1.25) doit se faire avec grande probabilité, uniformément en δ, vu que le module de continuité

$$\sup_{\{s, P(Ks-Ks_M)\leq\delta\}} |(P-P_n)(Ks-Ks_M)|$$

est aléatoire. Ceci est possible, par exemple en contrôlant les fluctuations du module de continuité autour de sa moyenne à δ fixé, puis en utilisant la croissance en δ du supremum pour passer à l'uniformité en cette variable.

Une notion centrale dans l'étude des vitesses rapides est la relation dite de marge qu'il peut exister entre la variance et l'excès de risque des fonctions indexantes. Cette relation a été formulée pour la première fois par Mammen et Tsybakov [60] (1999) dans le cadre de l'analyse discriminante. Par la suite, Tsybakov [75] a appliqué cette notion au contrôle de l'excès de

risque en classification binaire, ainsi que de nombreux auteurs en théorie de l'apprentissage statistique.
En suivant Massart et Nédélec [62], une relation de marge pour un ensemble de fonctions $f \in \mathcal{F}$ d'espérances $Pf \geq 0$, s'écrit

$$\sqrt{\mathrm{Var}\left(f\right)} \leq w\left(\sqrt{Pf}\right) \ , \tag{1.28}$$

où w est une fonction de $\mathbb{R}_+$ dans $\mathbb{R}_+$ croissante et continue telle que $x \longrightarrow w\left(x\right)/x$ est décroissante sur $\mathbb{R}_+^*$ avec $w\left(1\right) \geq 1$. Les exemples que nous connaissons dans les contextes tels que la classification binaire, l'estimation d'un ensemble de niveau, l'estimation de la densité ou la régression sont généralement réglés par des fonctions puissance :

$$\mathrm{Var}\left(Ks - Ks_M\right) \leq \kappa * \left(P\left(Ks - Ks_M\right)\right)^{\beta} \ , \tag{1.29}$$

où $\kappa > 0$ et $\beta \in \left]0, 1\right]$. La variance des fonctions indexantes apparaissant naturellement dans le contrôle du module de continuité du processus empirique, les relations du type (1.29) permettent en effet d'atteindre (1.25), où l'on demande un contrôle, non pas par la variance, mais par l'excès de risque des fonctions considérées.
Donnons un exemple de relation de marge, dans le contexte de l'estimation d'un excès de masse. Cet exemple est en fait une retraduction inédite des hypothèses de Polonik [65].

Exemple (estimation par excès de masse) *Soit ξ une variable aléatoire de loi P sur un ensemble mesurable $(\mathcal{Z}, \mathcal{T})$ et μ une mesure de référence sur cet ensemble. Par exemple, si $\mathcal{Z} = \mathbb{R}^n$, on prend généralement pour μ la mesure de Lebesgue sur $\mathcal{Z}$. Soit $\lambda > 0$, on cherche à estimer l'ensemble C_λ d'excès de masse de niveau λ,*

$$C_\lambda = \arg\max_{C \in \mathcal{T}} \left\{P\left(C\right) - \lambda\mu\left(C\right)\right\} \ . \tag{1.30}$$

Dans le cas où P admet une densité f par rapport à la mesure μ, on retrouve facilement que

$$C_\lambda = \left\{z \in \mathcal{Z} \text{ t.q. } f\left(z\right) \geq \lambda\right\}$$

est l'ensemble des points de niveau supérieur ou égal à λ pour la densité f. La formulation (1.30) permet d'envisager ce problème du point de vue de la M-estimation. Le fait que le problème d'optimisation se définisse sur des ensembles plutôt que sur des fonctions n'est nullement une gêne puisque, pour se ramener au cadre fonctionnel, il suffit de considérer les indicatrices d'ensembles. Le contraste s'écrit alors

$$K : C \in \mathcal{T} \mapsto \left(\left(KC\right) : z \in \mathcal{Z} \mapsto \lambda\mu\left(C\right) - 1_C\left(z\right)\right)$$

et on retrouve ainsi la formulation standard

$$C_\lambda = \arg\min_{C \in \mathcal{T}} P\left(KC\right) \ .$$

On définit alors l'excès de risque en un point C,

$$\begin{aligned}
\ell\left(C_\lambda, C\right) &= P\left(KC - KC_\lambda\right) \\
&= P\left(\lambda\mu\left(C\right) - 1_C - \left(\lambda\mu\left(C_\lambda\right) - 1_{C_\lambda}\right)\right) \\
&= \left(P\left(C_\lambda\right) - \lambda\mu\left(C_\lambda\right)\right) - \left(P\left(C\right) - \lambda\mu\left(C\right)\right) \geq 0 \ ,
\end{aligned}$$

et, si P a une densité f par rapport à μ, alors

$$\begin{aligned}
\ell\left(C_\lambda, C\right) = P\left(KC\right) - P\left(KC_\lambda\right) &= \int_{C_\lambda} \left(f - \lambda\right) d\mu - \int_{C} \left(f - \lambda\right) d\mu \\
&= \int_{C_\lambda \setminus C} \left(f - \lambda\right) d\mu - \int_{C \setminus C_\lambda} \left(f - \lambda\right) d\mu \\
&= \int_{C \Delta C_\lambda} \left|f - \lambda\right| d\mu \ ,
\end{aligned}$$

où $A\Delta B := (A\backslash B) \cup (A\backslash B)$ *est la différence symmétrique entre les ensembles* A *et* B*. On se place désormais dans ce cas et on suppose que la densité* f *est uniformément bornée sur* $\mathcal{Z}$ *par une constante* $B > 0$*. Alors pour tout* $\eta > 0$*, si* $|f(z) - \lambda| > \eta$ *pour un certain* $z \in \mathcal{Z}$*, on a* $\frac{B}{\eta} |f(z) - \lambda| > B \geq f$*, donc*

$$\begin{aligned} P(C\Delta C_\lambda) &\leq \frac{B}{\eta} \int_{C\Delta C_\lambda} |f - \lambda| \, d\mu + P(|f - \lambda| \leq \eta) \\ &\leq \frac{B}{\eta} \ell(C_\lambda, C) + P(|f - \lambda| \leq \eta) \ . \end{aligned}$$

D'où, si l'on suppose maintenant qu'il existe $c, \gamma > 0$ *tels que*

$$P(|f - \lambda| \leq \eta) \leq c\eta^\gamma$$

pour tout $\eta > 0$*, on obtient*

$$P(C\Delta C_\lambda) \leq \inf_{\eta > 0} \left\{ \frac{B}{\eta} \ell(C, C_\lambda) + c\eta^\gamma \right\}$$

et pour $\eta = \left(\frac{B}{c\gamma} \ell(C, C_\lambda) \right)^{\frac{1}{\gamma+1}} > 0$ *on trouve*

$$P(C\Delta C_\lambda) \leq \kappa * \ell(C, C_\lambda)^{\frac{\gamma}{\gamma+1}} \ ,$$

avec κ *une constante positive. Finalement, en remarquant*

$$\begin{aligned} \mathrm{Var}(KC - KC_\lambda) &= \mathrm{Var}(\lambda\mu(C\Delta C_\lambda) - 1_{C\Delta C_\lambda}) \\ &= \mathrm{Var}(1_{C\Delta C_\lambda}) = P(C\Delta C_\lambda) \end{aligned}$$

on obtient une relation de marge donnée par

$$\mathrm{Var}(KC - KC_\lambda) \leq \kappa * \ell(C, C_\lambda)^{\frac{\gamma}{\gamma+1}} \ .$$

On peut ainsi revisiter les résultats de Polonik [65] par le formalisme des relations de marge et on vérifie aisément qu'une application directe des résultats généraux de bornes supérieures d'excès de risque par relation de marge, comme par exemple le théorème principal (et non asymptotique) de l'article de Massart et Nédélec [62], permet de retrouver les vitesses (asymptotiques) données par Polonik dans le cadre de l'estimation par excès de masse.

Nous clôturons cette section, dédiée au contrôle général de l'excès de risque d'un M-estimateur, en donnant un résultat récent de Massart et Nédélec, retranscrit ici avec nos notations. Nous ferons appel dans le théorème qui suit à la définition suivante :

Définition On note $\mathcal{C}_1$ l'ensemble des fonctions ψ de $\mathbb{R}_+$ dans $\mathbb{R}_+$, croissantes et continues, telles que $x \longrightarrow \psi(x)/x$ est décroissante sur $\mathbb{R}_+^*$ avec $\psi(1) \geq 1$.

Théorème (Massart et Nédélec, [62]) Soit $(K, \mathcal{S}, P)$ un triplet tel que $K : S \longrightarrow L_1^-(P)$ soit un contraste. On se donne un modèle $M \subset \mathcal{S}$. Soit d une pseudo-distance sur $\mathcal{S} \times \mathcal{S}$ telle que

$$\forall s \in \mathcal{S}, \ \mathrm{Var}(Ks - Ks_*) \leq d^2(s_*, s) \ , \tag{1.31}$$

et soient ϕ et w des éléments de $\mathcal{C}_1$. On suppose que

$$\forall s \in \mathcal{S}, \ d(s_*, s) \leq w\left(\sqrt{\ell(s_*, s)} \right) \tag{1.32}$$

et que, pour tout $t \in M$, pour tout $\sigma > 0$ tel que $\phi(\sigma) \leq \sqrt{n}\sigma^2$,

$$\sqrt{n}\mathbb{E}\left[\sup_{s\in M, d(t,s)\leq\sigma}\{(P_n - P)(Kt - Ks)\}\right] \leq \phi(\sigma) \ . \tag{1.33}$$

Soit ε_* l'unique solution de l'équation

$$\sqrt{n}\varepsilon_*^2 = \phi(w(\varepsilon_*)) \ . \tag{1.34}$$

Il existe alors une constante absolue $\kappa > 0$ telle que pour tout $y \geq 0$, on a

$$\mathbb{P}\left[\ell(s_*, s_n(M)) > 2\ell(s_*, s_M) + \kappa\left(\varepsilon_*^2 + \frac{\left(1 \wedge w^2(\varepsilon_*)\right)}{n\varepsilon_*^2}y\right)\right] \leq \exp(-y) \ . \tag{1.35}$$

En particulier, on a le contrôle en moyenne suivant,

$$\mathbb{E}\left[\ell(s_*, s_n(M))\right] \leq 2\left(\ell(s_*, s_M) + \kappa\varepsilon_*^2\right) \ . \tag{1.36}$$

Avant commenter plus précisement le théorème précédent, notons que l'on a légèrement simplifié ici le résultat initial de Massart et Nédélec (Theorem 2, [62]) où les auteurs considèrent le cas plus général de l'excès de risque de ρ-minimiseurs empiriques sur M - voir la définition (1.4) ci-dessus - plutôt que du seul M-estimateur $s_n(M)$; ce qui permet en particulier de se libérer du problème de l'existence d'un tel estimateur. Cette généralisation ne change que très peu la forme du résultat et les arguments de preuve développés dans [62].
On distingue dans le théorème précédent deux types d'hypothèses. En effet, les conditions (1.31) et (1.32) permettent tout d'abord de retrouver la condition de marge exprimée plus généralement en (1.28), et mettent donc en relation la variance des fonctions d'intérêt avec leur excès de risque. En second lieu, le contrôle requis en (1.33) concerne la complexité du modèle M considéré, qui est ici exprimée *via* la fonction ϕ. En effet, on vérifie aisément que la quantité à gauche de l'inégalité (1.33) est croissante pour l'inclusion des modèles, une propriété légitimement attendue pour une notion convenable de complexité. Dans un cas précis d'estimation (voir par exemple le cas des images binaires traité dans [62]), on calculera typiquement la fonction ϕ par des arguments de chaînage, faisant naturellement appel à des quantités telles que les entropies métriques, aléatoires ou à crochet, ou encore la dimension de Vapnik et Červonenkis.
Concernant les bornes d'excès de risque données en (1.35) et (1.36), on remarque que les quantités qui gèrent la concentration stochastique de l'excès de risque sont fonction de ε_*, qui est solution de l'équation de point fixe (1.34). Cet argument de point fixe est en essence tout à fait similaire à la résolution de l'inéquation (1.27), dont on peut rappeler qu'elle est sensée fournir une majoration fine de l'excès de risque. Le résultat de Massart et Nédélec confirme donc bien l'acuité de l'approche que nous avons présenté précédemment dans cette section.
Cependant, on remarque que les bornes (1.35) et (1.36) sont données à une constante multiplicative près. Ceci est dû en particulier aux arguments de chaînage utilisés dans la preuve de (1.35). L'approche générale présentée ici pose donc problème en ce qui concerne la validation théorique de l'heuristique de pente, où l'on demande un contrôle à la constante près des excès de risque théoriques et empiriques à modèle fixé. De plus, l'inéquation (1.26), qui est le point de départ de la démonstration donnée par Massart et Nédélec, permet seulement de considérer des bornes majorantes pour les quantités en jeu, et ne permet pas d'atteindre des minorations, nécessaires pour discuter la validité du phénomène de la pente. Nous abordons en détails ces problèmes dans la section suivante.

1.4 Bornes optimales pour les excès de risques à modèle fixé, dans le cas d'un contraste régulier

On pose ici la question de l'équivalence entre l'excès de risque et l'excès de risque empirique à modèle fixé,

$$P\left(Ks_n\left(M\right) - Ks_M\right) \sim P_n\left(Ks_M - Ks_n\left(M\right)\right) \ ?$$

Comme annoncé précédemment, ce fait présumé est la pierre angulaire de l'heuristique de pente. Afin d'acquérir un résultat aussi général que possible, on développe une méthodologie de preuve inédite, basée sur la notion de contraste régulier définie en Section 1.2. Le but est donc d'obtenir des bornes inférieures et supérieures en déviation, pour l'excès de risque et sa contrepartie empirique, suffisamment précises pour fournir un équivalent asymptotique de ces quantités par rapport au nombre de données n. On s'attachera dans la suite à éclairer cette question en donnant les grandes lignes de la preuve générale développée au chapitre 7. Nous donnons aussi, en fin de section, un exemple plus précis de bornes obtenues en régression hétéroscédastique sur des modèles de polynômes par morceaux.

On commence par réécrire les problèmes de bornes inférieures et supérieures de déviation, pour l'excès de risque et l'excès de risque empirique. Soit donc C et α deux quantités strictement positives. La question de la majoration de l'excès de risque avec grande probabilité s'exprime comme suit : trouver, à $\alpha > 0$ fixé, le plus petit $C > 0$ tel que

$$\mathbb{P}\left[P\left(Ks_n\left(M\right) - Ks_M\right) > C\right] \leq n^{-\alpha} \ .$$

On écrit alors, par définition du M-estimateur $s_n\left(M\right)$ comme minimiseur du risque empirique sur le modèle M,

$$\begin{aligned} & \mathbb{P}\left[P\left(Ks_n\left(M\right) - Ks_M\right) > C\right] \\ \leq \ & \mathbb{P}\left[\inf_{s \in M_C} P_n\left(Ks - Ks_M\right) \geq \inf_{s \in M_{>C}} P_n\left(Ks - Ks_M\right)\right] \\ = \ & \mathbb{P}\left[\sup_{s \in M_C} P_n\left(Ks_M - Ks\right) \leq \sup_{s \in M_{>C}} P_n\left(Ks_M - Ks\right)\right] \ , \end{aligned} \tag{1.37}$$

où

$$M_C := \left\{s \in M \ ; \ P\left(Ks_n\left(M\right) - Ks_M\right) \leq C\right\}$$

et

$$M_{>C} := M \backslash M_C = \left\{s \in M \ ; \ P\left(Ks_n\left(M\right) - Ks_M\right) > C\right\} \ .$$

De même, on peut réécrire la question de la minoration de l'excès de risque avec grande probabilité. On veut en effet cette fois trouver le plus grand $C > 0$ tel que

$$\mathbb{P}\left[P\left(Ks_n\left(M\right) - Ks_M\right) \leq C\right] \leq n^{-\alpha} \ .$$

On a alors, par définition du M-estimateur $s_n\left(M\right)$,

$$\begin{aligned} & \mathbb{P}\left[P\left(Ks_n\left(M\right) - Ks_M\right) \leq C\right] \\ \leq \ & \mathbb{P}\left[\inf_{s \in M_C} P_n\left(Ks - Ks_M\right) \leq \inf_{s \in M_{>C}} P_n\left(Ks - Ks_M\right)\right] \\ = \ & \mathbb{P}\left[\sup_{s \in M_C} P_n\left(Ks_M - Ks\right) \geq \sup_{s \in M_{>C}} P_n\left(Ks_M - Ks\right)\right] \ . \end{aligned} \tag{1.38}$$

Les formulations acquises en (1.37) et (1.38) sont en essence très proches des calculs menés par Bartlett et Mendelson dans [19], et permettent de ramener l'étude des bornes inférieures et supérieures pour l'excès de risque à la comparaison de deux quantités d'intérêt,

$$\sup_{s \in M_C} P_n\left(Ks_M - Ks\right) \quad \text{et} \quad \sup_{s \in M_{>C}} P_n\left(Ks_M - Ks\right) \ .$$

On écrit de plus, en posant $\mathcal{D}_L = \{s \in M \ ; \ P(Ks_n(M) - Ks_M) = L\}$,

$$\sup_{s \in M_C} P_n(Ks_M - Ks) = \sup_{0 \leq L \leq C} \left\{ \sup_{s \in \mathcal{D}_L} P_n(Ks_M - Ks) \right\}$$

et

$$\sup_{s \in M_{>C}} P_n(Ks_M - Ks) = \sup_{L > C} \left\{ \sup_{s \in \mathcal{D}_L} P_n(Ks_M - Ks) \right\} \ .$$

L'étude de l'excès de risque à modèle fixé se réduit donc au contrôle, à la constante près, de la quantité générique suivante,

$$\sup_{s \in \mathcal{D}_L} P_n(Ks_M - Ks) \ , L \geq 0 \ . \tag{1.39}$$

De manière très similaire, l'excès de risque empirique s'écrit, par définition du M-estimateur $s_n(M)$,

$$\begin{aligned} P_n(Ks_M - Ks_n(M)) &= \sup_{s \in M} P_n(Ks_M - Ks) \\ &= \sup_{L \geq 0} \left\{ \sup_{s \in \mathcal{D}_L} P_n(Ks_M - Ks) \right\} \ . \end{aligned} \tag{1.40}$$

Finalement, l'étude de l'excès de risque empirique se ramène encore au contrôle de la quantité donnée en (1.39). On a alors

$$\begin{aligned} \sup_{s \in \mathcal{D}_L} P_n(Ks_M - Ks) &= \sup_{s \in \mathcal{D}_L} \{(P_n - P)(Ks_M - Ks) + P(Ks_M - Ks)\} \\ &= \sup_{s \in \mathcal{D}_L} \{(P_n - P)(Ks_M - Ks)\} - L \ . \end{aligned} \tag{1.41}$$

Par la formule (1.41), il suffit d'étudier les suprema du processus empirique indexé par des fonctions contrastées, recentrées par la cible contrastée et d'excès de risque constant - égal à L - sur le modèle M.

Remarque : *Dans le cas où $s_n(M)$ est unique et où*

$$\forall C \geq 0, \ \sup_{s \in \mathcal{D}_C} P_n(Ks_M - Ks) \ \textit{est atteint} \ \left(= \max_{s \in \mathcal{D}_C} P_n(Ks_M - Ks) \right) \ ,$$

on a - par le même type de raisonnement qu'en (1.37) et (1.38) - la formule exacte suivante,

$$P(Ks_n(M) - Ks_M) = \arg\max_{L \geq 0} \left\{ \max_{s \in \mathcal{D}_L} P_n(Ks_M - Ks) \right\} \ . \tag{1.42}$$

Par (1.40) on a aussi la formule suivante,

$$P_n(Ks_M - Ks_n(M)) = \max_{L \geq 0} \left\{ \max_{s \in \mathcal{D}_L} P_n(Ks_M - Ks) \right\} \ . \tag{1.43}$$

Les formules (1.42) et (1.43) montrent que l'excès de risque et l'excès de risque empirique sur un modèle fixé ne sont pas de même nature, car le premier prend ses valeurs dans les arguments de la fonction $\Psi_n : L(\geq 0) \mapsto \max_{s \in \mathcal{D}_L} P_n(Ks_M - Ks)$, alors que le second se mesure d'après les images de la fonction Ψ_n. L'équivalence de l'excès de risque et de l'excès de risque empirique, si elle est vérifiée, serait donc un fait non trivial en général, émanant de l'équation de point fixe suivante,

$$\arg\max_{\mathbb{R}_+} \{\Psi_n\} \sim \max_{\mathbb{R}_+} \{\Psi_n\} \ .$$

Dans le but d'étudier les suprema du processus empirique apparaissant dans (1.41), on note que, sous de bonnes hypothèses, on dispose d'inégalités de concentration à droite (Bousquet, [28]) et à gauche (Klein [42], Klein et Rio [41]), de ces suprema autour de leur moyenne. On peut ainsi s'attendre à ce que les déviations de ces quantités soient, sous des hypothèses standards, négligeables devant leur moyenne,

$$\sup_{s\in\mathcal{D}_L}\left\{(P_n-P)\left(Ks_M-Ks\right)\right\}\sim\mathbb{E}\left[\sup_{s\in\mathcal{D}_L}\left\{(P-P_n)\left(Ks-Ks_M\right)\right\}\right]\ . \tag{1.44}$$

Il resterait alors à établir un contrôle à la constante près du terme de droite dans l'équivalence (1.44). On obtient un tel contrôle dans le cas des contrastes réguliers. En effet, on rappelle que si K est un contraste régulier pour le modèle M et la loi P (cf. Section 1.2), alors on a, pour tout $s\in M$,

$$Ks-Ks_M=\psi_0^s+\psi_{1,M}\cdot\left(s-s_M\right)+\psi_2\left(\psi_{3,M}\cdot\left(s-s_M\right)\right)\ , \tag{1.45}$$

avec ψ_0^s constante sur $\mathcal{Z}$ pour tout $s\in M$. On injecte alors (1.45) dans le terme de droite dans l'identité (1.44), et on obtient

$$\begin{aligned}&\mathbb{E}\left[\sup_{s\in\mathcal{D}_L}\left\{(P-P_n)\left(Ks-Ks_M\right)\right\}\right]\\ =\ &\underbrace{\mathbb{E}\left[\sup_{s\in\mathcal{D}_L}\left\{(P-P_n)\left(\psi_{1,M}\cdot\left(s-s_M\right)\right)\right\}\right]}_{\text{partie principale}}+\underbrace{\mathbb{E}\left[\sup_{s\in\mathcal{D}_L}\left\{(P-P_n)\left(\psi_2\left(\psi_{3,M}\cdot\left(s-s_M\right)\right)\right)\right\}\right]}_{\text{reste négligeable}}\ .\end{aligned}$$

Pour montrer maintenant que le terme de reste est bien négligeable devant la partie principale, il est nécessaire de faire appel à l'hypothèse de convergence en norme infinie du M-estimateur $s_n\left(M\right)$. En effet, le comportement de la fonction ψ_2 est, par hypothèse, typiquement quadratique (cf. Section 1.2). Donc, si les fonctions s considérées en argument du processus empirique $(P-P_n)\left(\psi_2\left(\psi_{3,M}\cdot\left(s-s_M\right)\right)\right)$ sont suffisamment proches de la projection s_M en norme infinie, il sera possible d'utiliser un principe de contraction, dû à Talagrand (voir [53], et aussi Theorem 7.4, Chapitre 7), et de négliger le reste devant la partie principale. L'hypothèse de consistance en norme infinie du M-estimateur considéré intervient donc pour contrôler suffisamment finement la norme infinie des arguments du processus empirique, en remplaçant dans les calculs, avec grande probabilité, le modèle M par la boule en norme infinie,

$$B_{L_\infty}\left(s_M,R_{M,n,\alpha}\right)=\left\{s\in M\ ;\ \left\|s-s_M\right\|_\infty\leq R_{M,n,\alpha}\right\}$$

où, par hypothèse de consistance en norme infinie, on a pour tout $n\geq n_1$,

$$\mathbb{P}\left(\left\|s_n\left(M\right)-s_M\right\|_\infty\leq R_{M,n,\alpha}\right)\geq 1-n^{-\alpha}\text{ et }R_{M,n,\alpha}\leq\frac{A_{cons}}{\left(\ln n\right)^{1/4}}\ , \tag{1.46}$$

pour une constante absolue $A_{cons}>0$ et un entier naturel n_1. La couche $\mathcal{D}_L$ du modèle M, utilisée dans les raisonnements précédents, n'est donc pas exactement celle utilisée dans les preuves, car en vue du contrôle du terme de reste, on raffine le raisonnement et on considère en fait la couche

$$\tilde{\mathcal{D}}_L=\mathcal{D}_L\cap B_{L_\infty}\left(s_M,R_{M,n,\alpha}\right)\ ,$$

localisée à la fois en excès de risque et en norme infinie. L'hypothèse (1.46) permet alors d'appliquer un principe de contraction qui fournit

$$\mathbb{E}\left[\sup_{s\in\tilde{\mathcal{D}}_L}\left\{(P-P_n)\left(Ks-Ks_M\right)\right\}\right]\sim\mathbb{E}\left[\sup_{s\in\tilde{\mathcal{D}}_L}\left\{(P-P_n)\left(\psi_{1,M}\cdot\left(s-s_M\right)\right)\right\}\right]\ . \tag{1.47}$$

Bien que l'encadrement à la constante près du terme de droite dans l'équivalence (1.47) soit une partie particulièrement technique de la preuve exposée au chapitre 7 - et aussi dans les preuves des cas plus restreints des chapitres qui le précèdent -, il est néamoins possible de se faire assez simplement une idée de son équivalent asymptotique. On admet pour cela que sous une hypothèse de structure de l'espace vectoriel M_0, supposé de dimension finie D_M et sous-jacent à l'espace affine $\mathrm{Aff}(M)$, on obtient

$$\mathbb{E}\left[\sup_{s\in\tilde{\mathcal{D}}_L}\left\{(P-P_n)\left(\psi_{1,M}\cdot(s-s_M)\right)\right\}\right]\sim\mathbb{E}\left[\sup_{s\in\mathcal{D}_L}\left\{(P-P_n)\left(\psi_{1,M}\cdot(s-s_M)\right)\right\}\right]\ . \tag{1.48}$$

On revient donc au cas - plus simple - de la couche $\mathcal{D}_L$, l'équivalence (1.48) étant justifiée pour des modèles M_0 dits "à base localisée". L'hypothèse de base localisée (voir Massart [61], Section 7.4.2, et aussi Chapitre 7, Section 7.1.3) est une hypothèse classique en sélection de modèles, qui stipule un contrôle de la norme infinie des fonctions d'un espace hermitien, par la norme infinie des coefficients de ces fonctions dans une base orthonormale pour le produit scalaire considéré. Cette hypothèse est typiquement vérifiée pour des histogrammes ou des polynômes par morceaux définis sur des partitions régulières, et aussi pour des développements en ondelettes à support compact (voir Massart [61], Section 7.4.2).

D'après la définition d'un contraste régulier donnée à la Section 1.2, on a, pour toute fonction $s\in M$ telle que $\|s-s_M\|_\infty\leq\delta\leq L_H^{-1}$,

$$(1-L_H\delta)\,\|s-s_M\|_{H,M}^2\leq P\left(Ks-Ks_M\right)\leq(1+L_H\delta)\,\|s-s_M\|_{H,M}^2\ . \tag{1.49}$$

Ainsi, l'encadrement (1.49) suppose l'équivalence de l'excès de risque de toute fonction suffisamment proche en norme infinie de la projection s_M, avec la norme hilbertienne $\|\cdot\|_{H,M}$. On s'attend donc à avoir, pour les couches d'intérêt,

$$\mathbb{E}\left[\sup_{s\in\mathcal{D}_L}\left\{(P-P_n)\left(\psi_{1,M}\cdot(s-s_M)\right)\right\}\right]\sim\mathbb{E}\left[\sup_{\left\{s\in M;\|s-s_M\|_{H,M}=L\right\}}\left\{(P-P_n)\left(\psi_{1,M}\cdot(s-s_M)\right)\right\}\right]\ .$$

De plus, en contrôlant la variance du supremum du processus empirique pour les fonctions considérées (voir Corollary 7.2, Chapitre 7), on aura aussi

$$\begin{aligned}&\mathbb{E}\left[\sup_{\left\{s\in M;\|s-s_M\|_{H,M}=L\right\}}\left\{(P-P_n)\left(\psi_{1,M}\cdot(s-s_M)\right)\right\}\right]\\ \sim\ &\mathbb{E}^{1/2}\left[\left(\sup_{\left\{s\in M;\|s-s_M\|_{H,M}=L\right\}}\left\{(P-P_n)\left(\psi_{1,M}\cdot(s-s_M)\right)\right\}\right)^2\right]\\ &\underset{\text{inégalité de Cauchy-Schwarz}}{=}\sqrt{\frac{L\cdot\sum_{k=1}^{D_M}\mathrm{Var}\left(\psi_{1,M}\cdot\varphi_k\right)}{n}}\ ,\end{aligned} \tag{1.50}$$

où $(\varphi_k)_{k=1}^{D_M}$ est une base orthonormale quelconque de $\left(M,\|\cdot\|_{H,M}\right)$.

Finalement, pour obtenir des développements asymptotiques d'ordre 1 pour l'excès de risque et l'excès de risque empirique, il suffit d'injecter l'équivalent (1.50) dans les calculs initiaux.

On obtient ainsi pour l'excès de risque,

$$\begin{aligned} P\left(Ks_n\left(M\right)-Ks_M\right) &\sim \arg\max_{L\geq 0}\left\{\max_{s\in\mathcal{D}_L} P_n\left(Ks_M-Ks\right)\right\} \\ &\sim \arg\max_{L\geq 0}\left\{\max_{s\in\mathcal{D}_L} \mathbb{E}\left[\left(P_n-P\right)\left(Ks_M-Ks\right)\right]-L\right\} \\ &\sim \arg\max_{L\geq 0}\left\{\sqrt{\frac{L\cdot\sum_{k=1}^{D_M}\mathrm{Var}\left(\psi_{1,M}\cdot\varphi_k\right)}{n}}-L\right\} \\ &= \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\ , \end{aligned}$$

où $\mathcal{K}_{1,M}^2=\frac{1}{D_M}\sum_{k=1}^{D_M}\mathrm{Var}\left(\psi_{1,M}\cdot\varphi_k\right)$ sera une quantité - indépendante de la base $(\varphi_k)_{k=1}^{D_M}$ choisie - typiquement encadrée par deux constantes absolues strictement positives et qui dépendent des conditions, suffisamment bonnes, du problème considéré. Enfin, concernant le contrôle de l'excès de risque empirique, on obtient,

$$\begin{aligned} P_n\left(Ks_M-Ks_n\left(M\right)\right) &= \max_{L\geq 0}\left\{\max_{s\in\mathcal{D}_L} P_n\left(Ks_M-Ks\right)\right\} \\ &\sim \max_{L\geq 0}\left\{\sqrt{\frac{L\cdot\sum_{k=1}^{D_M}\mathrm{Var}\left(\psi_{1,M}\cdot\varphi_k\right)}{n}}-L\right\} \\ &= \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\ . \end{aligned}$$

L'équivalence

$$P\left(Ks_n\left(M\right)-Ks_M\right)\sim P_n\left(Ks_M-Ks_n\left(M\right)\right)\left(\sim\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\right)$$

est ainsi justifiée.

Nous terminons cette section par un exemple de résultat précis acquis au Chapitre 3. En effet, dans le cas des polynômes par morceaux, en régression hétéroscédatique bornée avec *design* aléatoire, on obtient le résultat suivant en considérant l'estimateur des moindres carrés.
Soit Leb la mesure de Lebesgue sur $[0,1]$ et soit $\alpha>0$. On suppose que $\mathcal{X}=[0,1]$ et que P^X a une densité f par rapport à la mesure de Lebesgue Leb vérifiant, pour des constantes positives $c_{\min}$ et $c_{\max}$,

$$0<c_{\min}\leq f\left(x\right)\leq c_{\max}<+\infty,\ x\in[0,1]\ .$$

Soit M un modèle de polynômes par morceaux sur une partition finie $\mathcal{P}$ de $[0,1]$, de dégrés inférieurs ou égaux à r. On suppose de plus qu'il existe A_- et A_+ deux constantes positives telles que

$$A_-\left(\ln n\right)^2\leq D_M\leq A_+\frac{n}{\left(\ln n\right)^2}\ . \tag{1.51}$$

Alors, sous de bonnes hypothèses sur la partition considérée et sur le bruit, il existe une constante positive A_0, ne dépendant que des constantes du problème, telle qu'en notant

$$\varepsilon_n=A_0\max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4},\left(\frac{D_M\ln n}{n}\right)^{1/4}\right\}\ ,$$

il existe un entier n_0 ne dépendant que des constantes du problème, tel que pour tout $n\geq n_0$, on a

$$\mathbb{P}\left[\left(1+\varepsilon_n\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\geq P\left(Ks_n\left(M\right)-Ks_M\right)\geq\left(1-\varepsilon_n\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\right]\geq 1-10n^{-\alpha} \tag{1.52}$$

et

$$\mathbb{P}\left[\left(1+\varepsilon_n^2\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \geq P_n\left(Ks_M - Ks_n\left(M\right)\right) \geq \left(1-\varepsilon_n^2\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\right] \geq 1 - 5n^{-\alpha} \ . \quad (1.53)$$

D'après les encadrements (1.52) et (1.53), la partie principale du développement asymptotique des excès de risque et excès de risque empirique est bien

$$\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 = \frac{1}{4n}\sum_{k=1}^{D_M}\mathrm{Var}\left(\psi_{1,M}\cdot\varphi_k\right) \ , \quad (1.54)$$

et les excès de risque sont bien équivalents entre eux. Il est à noter que, bien que la dimension du modèle M considéré soit fixe, elle n'est pas considérée comme une constante du problème, et peut dépendre du nombre de données n, dans les limites permises en (1.51), cette dépendance devenant cruciale dès qu'il sera question ci-dessous de sélection de modèles du point de vue *non-asymptotique*. De plus, bien que sous le jeu d'hypothèses considéré pour établir ces résultats, le terme de complexité $\mathcal{K}_{1,M}^2$ se comporte typiquement comme une constante, au sens où il est encadré par deux constantes strictement positives qui dépendent des conditions du problème considéré, la partie principale des excès de risque *n'est pas*, en général, linéaire en la dimension D_M. Au contraire, lorsque le bruit est fortement hétéroscédastique, les excès de risque - et donc, comme nous le verrons, la pénalité idéale du problème de sélection associé - s'approchent mal par une fonction linéaire en la dimension, comme le démontre Arlot [6], dans le cas de la régression hétéroscédastique sur des modèles d'histogrammes. Par contre, lorsque l'on suppose que le bruit est homoscédastique,

$$\sigma\left(X\right) \equiv \sigma > 0 \quad p.s. \ ,$$

et que l'on choisit pour modèle M l'ensemble des histogrammes formés par une partition $\mathcal{P}$ de $\mathcal{X}$ à D_M éléments, alors on a

$$\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 = \sigma^2\frac{D_M}{n} + \frac{D_M}{n}\cdot\frac{1}{|\mathcal{P}|}\sum_{I\in\mathcal{P}}\mathbb{V}\left[\mathbb{E}\left[Y\,|X\right]|X\in I\right] \ . \quad (1.55)$$

En négligeant le second terme à droite de l'égalité (1.55), ce qui sera typiquement justifié pour une partition suffisamment grande -et bien adaptée -, on trouve donc un équivalent asymptotique égal à $\sigma^2 D_M/n$, c'est-à-dire la moitié de la pénalité de Mallows ([59]). En effet, à travers les sections suivantes, nous comprendrons que les encadrements (1.52) et (1.53) justifient en particulier dans ce cas, de manière non-asymptotique, l'emploi de la pénalité de Mallows pour traiter ce problème de régression homoscédastique, et qu'elles permettent plus généralement de valider l'heuristique de pente en régression hétéroscédastique.

1.5 Sélection de modèles

Comme nous l'avons expliqué aux Sections 1.3 et 1.4, la vitesse de convergence d'un M-estimateur à modèle fixé dépend de la complexité du modèle considéré. Pourtant, nous n'avons pas abordé un problème crucial : le choix du modèle M servant à définir le M-estimateur.
En effet, le cadre non-paramétrique dans lequel nous nous plaçons ne nous permet pas de supposer l'appartenance de la cible s_* d'estimation à un modèle M particulier et accessible en pratique, comme par exemple un modèle de dimension finie. Ainsi, l'idée naturelle de la sélection de modèles est de se donner une collection de modèles $\mathcal{M}_n$ et la collection de M-estimateurs associés $\{s_n\left(M\right),\ M\in\mathcal{M}_n\}$, le but étant alors de choisir le meilleur estimateur

au sens du risque. Moyennant alors un *a priori* sur la cible, comme par exemple une hypothèse de régularité, on peut choisir une collection de modèles $\mathcal{M}_n$ ayant de bonnes propriétés d'approximation pour la cible considérée, au sens du risque. Ceci permet alors de supposer que le terme de biais décroît lorsque la complexité du modèle augmente, et que l'excès de risque à modèle fixé (1.9) des M-estimateurs considérés croît avec cette complexité. Le but effectif de la procédure de sélection de modèles est alors de trouver le meilleur compromis "biais-variance"parmi les modèles considérés.
Soit donc $\mathcal{M}_n$ une collection finie de modèles dont le cardinal dépend du nombre de données n, soit $\{s_n(M),\ M \in \mathcal{M}_n\}$ la collection des M-estimateurs associés. Le but est de retrouver le meilleur estimateur parmi la collection que l'on s'est donnée, la cible de la procédure de sélection de modèles est donc

$$M_* \in \arg\min_{M\in\mathcal{M}_n} \{P(Ks_n(M))\} \ . \tag{1.56}$$

L'estimateur en le modèle cible $s_n(M_*)$ est appelé l'oracle. On remarque alors que le but n'est pas ici d'identifier un "vrai"modèle mais de retrouver l'estimateur ayant la meilleure qualité de prédiction au sens du risque.
Par ailleurs, d'après l'égalité (1.56), on a aussi

$$M_* \in \arg\min_{M\in\mathcal{M}_n} \{\ell(s_*, s_n(M))\} \ . \tag{1.57}$$

A partir de (1.56), on pourrait alors penser à remplacer, comme c'est le cas à modèle fixé, le risque sous la loi inconnue par sa version empirique et sélectionner

$$\widehat{M} \in \arg\min_{M\in\mathcal{M}_n} \{P_n(Ks_n(M))\} \ . \tag{1.58}$$

Toutefois, par définition des M-estimateurs, le risque empirique $P_n(Ks_n(M))$ décroît le long d'une suite croissante de modèles et en pratique la procédure définie en (1.58) va systématiquement choisir des modèles de forte complexité, négligeant ainsi le compromis biais-variance, et menant à des estimateurs de faible performance devant l'oracle. Autrement dit, la procédure (1.58) permet de choisir l'estimateur qui "s'adapte"le mieux aux données et non pas celui qui a la meilleure capacité de prédiction. Ceci vient du fait que le risque empirique sous-estime systématiquement le vrai risque. En effet, on peut écrire

$$P(Ks_n(M)) = PKs_M + \underbrace{P(Ks_n(M) - Ks_M)}_{\geq 0}$$

et

$$P_n(Ks_n(M)) = P_nKs_M - \underbrace{P_n(Ks_M - Ks_n(M))}_{\geq 0} \ .$$

Si l'on considère alors $PKs_M \approx P_nKs_M$ on a bien $P_n(Ks_n(M)) < P(Ks_n(M))$ et le biais entre le risque sous la vraie loi et le risque empirique est alors de l'ordre de

$$(P - P_n)(Ks_n(M) - Ks_M) \geq 0 \ . \tag{1.59}$$

De la même manière que le terme de variance dans la décomposition de l'excès de risque (1.10), la quantité (1.59) mesure la complexité du modèle M considéré. À la suite de ce constat, une idée naturelle est, pour atteindre l'oracle, de remplacer dans (1.56) le vrai risque qui nous est inconnu par le risque empirique, ajouté d'une quantité positive pen(M) appelée pénalité et sensée débiaiser le risque empirique par rapport au vrai risque.
Plus précisément, si l'on se donne une fonction pen : $\mathcal{M}_n \to \mathbb{R}_+$, on peut définir une procédure de sélection de modèles en choisissant

$$\widehat{M} \in \arg\min_{M\in\mathcal{M}_n} \{P_n(Ks_n(M)) + \text{pen}(M)\} \ , \tag{1.60}$$

et on procèdera ainsi à une sélection de modèles *par pénalisation*. L'idée d'ajouter au risque empirique une pénalité rendant compte de la complexité des modèles et permettant de débiaiser, au moins asymptotiquement, l'estimation du risque sur chaque modèle, remonte aux travaux fondateurs d'Akaike [1] et [2], et Mallows [59]. Depuis une quinzaine d'années, la recherche dans le domaine de la sélection de modèles s'est intensifiée. Tout d'abord, les travaux de Birgé et Massart [22], et Barron, Birgé et Massart [13] ont permis de dégager les principes généraux de la sélection de modèles par pénalisation, et en particulier le lien qu'il existe entre la qualité d'approximation des modèles considérés, émanant de la théorie de l'approximation, et les propriétés minimax des estimateurs sélectionnés, avec en conséquence leurs propriétés d'adaptation. Ces auteurs ont développé une théorie *non-asymptotique* de la sélection de modèles, par l'utilisation, centrale dans ce contexte, des inégalités de concentration de type Talagrand pour les suprema de processus empiriques. Ils ont ainsi pu mettre en évidence le rôle joué par la complexité de la collection de modèles $\mathcal{M}_n$ considérée dans le choix de la pénalité. À la suite de ces travaux, de nombreuses méthodes de pénalisation ont été proposées, comme par exemple les complexités de Rademacher (Koltchinskii [43], Bartlett et *al.* [16]), les complexités de Rademacher locales (Bartlett et *al.* [17], Koltchinskii [44]) ou encore les pénalités *bootstrap*, de rééchantillonnage et V-fold (Efron [38], Arlot [7] et [5]).
Pour mieux comprendre l'enjeu de ces stratégies, définissons la quantité suivante

$$\mathrm{pen}_{\mathrm{id}}(M) = (P - P_n)(Ks_n(M)) \tag{1.61}$$

appelée pénalité idéale sur le modèle M. On a alors, par (1.57),

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \{P_n K s_n(M) + \mathrm{pen}_{\mathrm{id}}(M)\} \ .$$

Afin de fournir des procédures générales performantes de sélection de modèles, il convient dès lors de s'attacher à comprendre le comportement de la pénalité idéale.
Dans le cas où le nombre de modèles considérés, c'est-à-dire le cardinal de $\mathcal{M}_n$, est polynômial en le nombre de données n, une bonne pénalité est nécessairement une pénalité dont l'ordre de grandeur est celui de $\mathrm{pen}_{\mathrm{id}}(M)$ pour chaque modèle M - à une constante près biensûr -, ou plutôt pour chaque modèle de dimension raisonnable. En effet, les inégalités de concentration pour le supremum du processus empirique, qui gèrent les déviations des quantités en jeu, sont sous-gaussiennes et permettent donc de sommer les probabilités de dévier autour de la moyenne pour des collections de modèles polynômiales. Par contre, si la complexité de la collection de modèles est plus forte, par exemple exponentielle, il faut contrecarrer les déviations en choisissant une pénalité plus élevée, typiquement d'un facteur logarithmique dépendant de cette complexité. Mais ceci sort du cadre de notre étude. Pour des théorèmes généraux prenant en compte la complexité de la collection de modèles considérée, on pourra consulter Massart [61].

1.6 Pénalités minimales et heuristique de pente

Un des buts de l'analyse non-asymptotique des procédures de sélection de modèles est de fournir des inégalités dites oracles, qui rendent compte de la performance de ces procédures à un nombre fixé de données. Il existe plusieurs variantes de telles inégalités ; nous nous intéressons ici aux inégalités oracles dites trajectorielles.
On dira qu'une procédure de sélection de modèles, définie par (1.60), satisfait une inégalité oracle trajectorielle si on a avec grande probabilité (par exemple $1 - Ln^{-2}$, où L est une constante)

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq C \inf_{M \in \mathcal{M}_n} \{\ell(s_*, s_n(M))\} \ , \tag{1.62}$$

où C est une constante positive. Ainsi, l'excès de risque de l'estimateur sélectionné est avec grande probabilité de l'ordre de celui de l'oracle. De plus, lorsque

$$C = 1 + \varepsilon, \text{ avec } \varepsilon \to 0 \text{ pour } n \to \infty ,$$

on dira que la procédure est asymptotiquement optimale, puisque l'on retrouve à l'infini la performance de l'oracle. En effet, dans ce cas on obtient

$$\mathbb{P}\left(\frac{\ell\left(s_*, s_n\left(\widehat{M}\right)\right)}{\inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n\left(M\right)\right)\right\}} \longrightarrow 1 \right) = 1 .$$

Il existe un lien direct entre le comportement d'une pénalité pen dans (1.60) vis-à-vis de la pénalité idéale pen_{id} définie en (1.61) et la performance de la procédure de sélection de modèles associée, décrite par l'inégalité oracle (1.62). En effet, on a d'après (1.57), (1.60) et (1.61), pour tout modèle $M \in \mathcal{M}_n$,

$$\begin{aligned}
\ell\left(s_*, s_n\left(\widehat{M}\right)\right) &= P_n K s_n\left(\widehat{M}\right) + \text{pen}_{\text{id}}\left(\widehat{M}\right) - PKs_* \\
&= P_n K s_n\left(\widehat{M}\right) + \text{pen}\left(\widehat{M}\right) + \left(\text{pen}_{\text{id}} - \text{pen}\right)\left(\widehat{M}\right) - PKs_* \\
&\leq P_n K s_n\left(M\right) + \text{pen}\left(M\right) + \left(\text{pen}_{\text{id}} - \text{pen}\right)\left(\widehat{M}\right) - PKs_* \\
&= \ell\left(s_*, s_n\left(M\right)\right) + \left(\text{pen} - \text{pen}_{\text{id}}\right)\left(M\right) + \left(\text{pen}_{\text{id}} - \text{pen}\right)\left(\widehat{M}\right).
\end{aligned}$$

Ainsi, en prenant l'infinimum sur les modèles considérés on a

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) + \left(\text{pen} - \text{pen}_{\text{id}}\right)\left(\widehat{M}\right) \leq \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n\left(M\right)\right) + \left(\text{pen} - \text{pen}_{\text{id}}\right)\left(M\right)\right\} . \quad (1.63)$$

Dès lors, si l'on a par exemple, uniformément sur $M \in \mathcal{M}_n$ et avec grande probabilité, l'encadrement

$$\text{pen}_{\text{id}}\left(M\right) \leq \text{pen}\left(M\right) \leq \text{pen}_{\text{id}}\left(M\right) + C * \ell\left(s_*, s_n\left(M\right)\right) \quad (1.64)$$

on obtient en utilisant (1.63) une inégalité oracle avec constante $C + 1$,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq (C+1) \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n\left(M\right)\right)\right\}. \quad (1.65)$$

De plus, si l'on a avec grande probabilité et pour tout modèle $M \in \mathcal{M}_n$,

$$\left|\left(\text{pen} - \text{pen}_{\text{id}}\right)\left(M\right)\right| \leq \varepsilon * \ell\left(s_*, s_n\left(M\right)\right)$$

où $\varepsilon < 1$, alors on obtient

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \frac{1+\varepsilon}{1-\varepsilon} \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n\left(M\right)\right)\right\}$$

et si $\varepsilon \to 0$ lorsque $n \to \infty$, alors on déduit de l'inégalité précédente l'optimalité asymptotique de la procédure.

L'enjeu de toute procédure de sélection par pénalisation étant de pénaliser correctement, observons les deux écueils possibles. Tout d'abord, on remarque qu'en surpénalisant, c'est-à-dire en faisant grandir la constante C dans le contrôle (1.64), la procédure de sélection de modèles satisfait encore une inégalité oracle donnée par (1.65); la constante dans cette inégalité étant moins bonne car égale à $C + 1$, mais asymptotiquement la performance de l'estimateur sélectionné reste de l'ordre de grandeur de l'oracle.

Par contre, sous-estimer la pénalité idéale peut avoir des conséquences dramatiques en termes de performance pour l'estimateur sélectionné. En effet, Birgé et Massart [23] ont montré dans le cas du modèle linéaire gaussien généralisé, l'existence d'une *pénalité minimale* $\text{pen}_{\min}$ vérifiant,

(F1) Si une pénalité $\mathrm{pen} : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ est telle que, pour tout modèle $M \in \mathcal{M}_n$,

$$\mathrm{pen}\,(M) \leq (1-\delta)\,\mathrm{pen}_{\min}\,(M)$$

avec $\delta > 0$, alors la procédure (1.60) sélectionne un modèle de très grande dimension et l'excès de risque de l'estimateur associé est très grand devant celui de l'oracle.

(F2) Si une pénalité $\mathrm{pen} : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ est telle que, pour tout modèle $M \in \mathcal{M}_n$,

$$\mathrm{pen}\,(M) \geq (1+\delta)\,\mathrm{pen}_{\min}\,(M)$$

avec $\delta > 0$, alors le modèle sélectionné est de dimension "raisonnable"et vérifie une inégalité oracle telle que (1.62) pour une contante $C > 1$.

(F3) Si $\mathrm{pen} \approx 2\,\mathrm{pen}_{\min}$ alors la procédure est quasiment optimale, au sens où elle vérifie une inégalité oracle avec constante proche de 1.

La conjonction des faits (**F1**), (**F2**) et (**F3**) constitue ce que les auteurs appellent l'*heuristique de pente*. Très récemment, Arlot et Massart [10] ont étendu la validité de l'heuristique de pente pour le cas de la régression avec un *design* aléatoire et un bruit hétéroscédastique sur des modèles par histogrammes, et ils ont identifié la forme générale de la pénalité minimale. Formellement, on peut écrire pour tout modèle $M \in \mathcal{M}_n$,

$$\begin{aligned}\ell\,(s_*, s_n\,(M)) &= P\,(Ks_n\,(M) - Ks_*)\\ &= P_n\,(Ks_n\,(M)) + P_n\,(Ks_M - Ks_n\,(M)) + (P_n - P)\,(Ks_* - Ks_M)\\ &\quad + P\,(Ks_n\,(M) - Ks_M) - P_n\,(Ks_*)\ .\end{aligned}$$

Donc, en posant

$$\begin{aligned}\mathrm{p}_1\,(M) &= P\,(Ks_n\,(M) - Ks_M)\\ \mathrm{p}_2\,(M) &= P_n\,(Ks_M - Ks_n\,(M))\\ \bar{\delta}\,(M) &= (P_n - P)\,(Ks_M - Ks_*)\end{aligned}$$

on obtient la décomposition suivante de l'excès de risque,

$$\ell\,(s_*, s_n\,(M)) = P_n\,(Ks_n\,(M)) + \mathrm{p}_1\,(M) + \mathrm{p}_2\,(M) - \bar{\delta}\,(M) - P_n\,(Ks_*)\,.$$

Pour une procédure de sélection définie en (1.60) on a alors, pour tout $M \in \mathcal{M}_n$,

$$\begin{aligned}&\ell\left(s_*, s_n\left(\widehat{M}\right)\right) - \left(\mathrm{p}_1\left(\widehat{M}\right) + \mathrm{p}_2\left(\widehat{M}\right) - \bar{\delta}\left(\widehat{M}\right) - \mathrm{pen}\left(\widehat{M}\right)\right)\\ &\leq \ell\,(s_*, s_n\,(M)) + \left(\mathrm{pen}\,(M) - \mathrm{p}_1\,(M) - \mathrm{p}_2\,(M) + \bar{\delta}\,(M)\right).\end{aligned}$$

Soit encore, en posant

$$\begin{aligned}\mathrm{pen}'_{\mathrm{id}}\,(M) &= \mathrm{pen}_{\mathrm{id}}\,(M) + (P_n - P)\,(Ks_*)\\ &= \mathrm{p}_1\,(M) + \mathrm{p}_2\,(M) - \bar{\delta}\,(M)\,,\end{aligned} \tag{1.66}$$

on obtient le contrôle suivant, pour tout $M \in \mathcal{M}_n$,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) - \left(\mathrm{pen}'_{\mathrm{id}}\left(\widehat{M}\right) - \mathrm{pen}\left(\widehat{M}\right)\right) \leq \ell\,(s_*, s_n\,(M)) + \left(\mathrm{pen}\,(M) - \mathrm{pen}'_{\mathrm{id}}\,(M)\right).$$

Arlot et Massart ont alors montré dans leur cas précis que la pénalité

$$\mathrm{pen}_{\min}\,(M) = \mathbb{E}\,[\mathrm{p}_2\,(M)] = \mathbb{E}\,[P_n\,(Ks_M - Ks_n\,(M))]$$

est une pénalité minimale au sens énoncé précédemment en (**F1-3**). Ainsi, en considérant que le terme centré $\bar{\delta}(M)$, présent dans l'expression de $\mathrm{pen}'_{\mathrm{id}}(M)$ donnée en (1.66), est négligeable devant le biais $P(Ks_M - Ks_*)$ du modèle M, le fait que la pénalité optimale soit approximativement égale à deux fois la pénalité minimale

$$\mathrm{pen}_{\mathrm{opt}}(M) \approx 2 * \mathrm{pen}_{\mathrm{min}}(M)$$

repose sur l'identité suivante,

$$\mathrm{p}_1(M) \sim \mathrm{p}_2(M)$$

c'est-à-dire

$$P(Ks_n(M) - Ks_M) \sim P_n(Ks_M - Ks_n(M)) \ . \tag{1.67}$$

Cette dernière relation exprime la proximité entre l'excès de risque sur le modèle M et l'excès de risque empirique, propriété qui peut être démontrée en établissant des bornes inférieures et supérieures fines avec grande probabilité pour ces deux quantités. Nous y parvenons dans le cadre des contrastes réguliers, comme expliqué en Section 1.4.
En pratique, l'heuristique de pente a une conséquence majeure pour la calibration optimale de la constante devant la pénalité. En effet, supposons connue la forme $\mathrm{pen}_{\mathrm{shape}} : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ de la pénalité idéale. Une méthode pour acquérir cette information est par exemple l'estimation de la pénalité idéale par une pénalité de type rééchantillonnage, voir à ce sujet Arlot [5], [7] et Lerasle [56]. Pour une certaine constante inconnue A_*, la pénalité $\left(A_* \cdot \mathrm{pen}_{\mathrm{shape}}\right)$ fournira donc une procédure quasiment optimale, au sens où elle vérifiera une inégalité oracle avec constante presque 1. Le but est alors de trouver $\hat{A}$ telle que $\left(\hat{A} \cdot \mathrm{pen}_{\mathrm{shape}}\right)$ soit quasiment optimale.
On suppose aussi connue une mesure de complexité D_M pour chaque modèle $M \in \mathcal{M}_n$. Typiquement, quand les modèles sont des espaces vectoriels de dimension finie, D_M est la dimension de M. Arlot et Massart [10] proposent alors l'algorithme suivant :

1. Calculer le modèle sélectionné $\widehat{M}(A)$ comme une fonction de $A > 0$,

$$\widehat{M}(A) \in \arg\min_{M \in \mathcal{M}_n} \left\{ P_n K(s_n(M)) + A\,\mathrm{pen}_{\mathrm{shape}}(M) \right\}.$$

2. Trouver $\hat{A}_{\mathrm{min}} > 0$ tel que $D_{\widehat{M}(A)}$ est "très grand" pour $A < \hat{A}_{\mathrm{min}}$ et "raisonnablement petit" pour $A > \hat{A}_{\mathrm{min}}$.

3. Sélectionner le modèle $\widehat{M} = \widehat{M}\left(2\hat{A}_{\mathrm{min}}\right)$.

Cet algorithme a déjà été appliqué avec succès dans des contextes variés tels que les modèles de mélanges (Maugis et Michel [63]), la classification non supervisée (Baudry [20]), l'estimation de modèles graphiques (Verzelen [86]), l'estimation de réserves pétrolières (Lepez [54]) et la génétique (Villers [87]).

1.7 Heuristique de pente pour l'estimation par minimum de contraste régulier

En utilisant les contrôles optimaux obtenus sur l'excès de risque et l'excès de risque empirique pour des modèles de dimension raisonnable, dans le cadre général de la M-estimation avec contraste régulier, on valide dans cette thèse l'heuristique de pente dans des situations classiques d'estimation non-paramétrique, tels que l'estimation de la densité par maximum de vraisemblance par sélection de modèles par histogrammes, ou encore la régression par moindres carrés par sélection de modèles linéaires.

Pour ce faire, nous adaptons, à partir des résultats nouveaux acquis à modèle fixé et décrits en Section 1.4, les preuves données par Arlot et Massart [10] dans le contexte de la régression hétéroscédastique par histogrammes. En effet, Arlot et Massart ont fourni une algèbre de preuve générale qui permet de valider l'heuristique de pente, sous de bonnes hypothèses suivant les problèmes considérés, dès lors que des contrôles optimaux sont acquis à modèle fixé, pour des complexités de modèles susceptibles d'être sélectionnées. En ce qui concerne l'établissement d'inégalités oracles, les grandes lignes de cette technique de preuve ont été esquissées en Section 1.6.
Au Chapitre 4, nous retrouvons ainsi les résultats obtenus par Arlot et Massart [10] dans le cas de la régression hétéroscédastique bornée, avec *design* aléatoire, sur des modèles linéaires d'histogrammes et nous généralisons ces résultats au cas des polynômes par morceaux. Nous donnons aussi un résultat plus structurel, qui permet de conclure que, si les modèles sont munis d'une base localisée, alors dès que les estimateurs des moindres carrés sont consistants en norme infinie vers les projetés de la cible qui leur sont respectivement associés, l'heuristique de pente est vérifiée, pour un certain jeu d'hypothèses sur la collection de modèles considérée.
Dans le cas de l'estimation de la densité par maximum de vraisemblance sur des modèles par histogrammes, on montre aussi que le critère AIC est asymptotiquement optimal, au sens où l'estimateur sélectionné vérifie une inégalité oracle non-asymptotique, avec une constante devant le risque de l'oracle qui tend vers un lorsque le nombre de données tend vers l'infini. De plus, on interprète la pénalité proposée par Akaike [2], comme deux fois la pénalité minimale introduite par Arlot et Massart [10]. C'est, à notre connaissance, le premier résultat validant l'heuristique de pente dans un cadre non-quadratique. Donnons, à titre d'exemple, la forme des résultats obtenus dans ce cadre auChapitre 5.

On montre que sous de bonnes hypothèses sur la collection de modèles et sur la cible s_*, si $\delta \in \left(0, \frac{1}{2}\right)$ et $L > 0$, et si l'on suppose qu'il existe un évènement de probabilité au moins $1 - A_p n^{-2}$ sur lequel, pour tout modèle $M \in \mathcal{M}_n$, tel que $D_M \geq A_{\mathcal{M},+} (\ln n)^2$, on a

$$(1-\delta)\frac{D_M - 1}{n} \leq \mathrm{pen}(M) \leq (1+\delta)\frac{D_M - 1}{n} ,$$

alors, pour un certain $\frac{1}{2} > \eta > (1-\beta_+)_+ /2$, il existe une constante A_3, un nombre entier n_0 ne dépendant que des constantes du problème et une suite

$$0 \leq \theta_n \leq \frac{L_{(\mathbf{SA})}}{(\ln n)^{1/4}}$$

pour une certaine certaine constante $L_{(\mathbf{SA})}$ ne dépendant de même que des constantes du problème, telles qu'avec probabilité au moins $1 - A_3 n^{-2}$, on a pour tout $n \geq n_0$,

$$D_{\widehat{M}} \leq n^{\eta+1/2} \tag{1.68}$$

et

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+2\delta}{1-2\delta} + \frac{4\theta_n}{(1-2\delta)^2}\right) \ell\left(s_*, s_n\left(M_*\right)\right) . \tag{1.69}$$

Les inégalités (1.68) et (1.69) fournissent ainsi une quantification précise, validant les points (**F2**) et (**F3**) de l'heuristique de pente exposée en Section 1.6.

De plus, s'il existe $A_{\mathrm{pen}} \in [0,1)$ et $A_p > 0$, tels que l'on a avec probabilité au moins $1 - A_p n^{-2}$, pour tout $M \in \mathcal{M}_n$,

$$0 \leq \mathrm{pen}(M) \leq A_{\mathrm{pen}} \frac{D_M - 1}{2n} ,$$

alors on montre qu'il existe deux constantes $A_1, A_2 > 0$ telles qu'avec probabilité $1 - A_1 n^{-2}$, on a pour tout $n \geq n_0 (A_{\text{pen}})$,

$$D_{\widehat{M}} \geq A_2 n \ln (n)^{-2} \tag{1.70}$$

et

$$\ell \left(s_*, s_n \left(\widehat{M} \right) \right) \geq \ln (n) \inf_{M \in \mathcal{M}_n} \left\{ \ell \left(s_*, s_n (M) \right) \right\} \ . \tag{1.71}$$

Les bornes inférieures décrites en (1.70) et (1.71) valident ainsi le point (**F1**) de l'heuristique de pente de Arlot, Birgé et Massart exposé en Section 1.6, et le phénomène de pente est ainsi démontré.

Conventions

- by a "constant", we always mean a quantity which does not depend on the number n of data. The linear dimension D or D_M of a model M is not treated as a constant, as it is allowed to depend on n.
- $L_{p_1,...,p_k}$, $L_{(\mathbf{A})}$ are generic constants, respectively depending on the constants $p_1, ..., p_k$ and on the constants appearing in the assumption set $(\mathbf{A})$.
- $n_0(p_1, ..., p_k)$, $n_0((\mathbf{A}))$ are generic positive integer-valued constants, respectively depending on the constants $p_1, ..., p_k$ and on the constants appearing in the assumption set $(\mathbf{A})$.
- we do not declare generic constants $\left(L_{p_1,...,p_k}, L_{(\mathbf{A})}, n_0(p_1, ..., p_k), n_0((\mathbf{A}))\right)$. For instance, the sentence "for all $n \geq n_0(p_1, ..., p_k)$, $a\left(n, L_{(\mathbf{A})}\right) \leq b\left(n, L_{\tilde{p}_1,...,\tilde{p}_k}\right)$" means that there exist a positive integer $n_0(p_1, ..., p_k)$ only depending on the constants $p_1, ..., p_k$ and two constants $L_{(\mathbf{A})}$, $L_{\tilde{p}_1,...,\tilde{p}_k}$ only depending on the assumption set $(\mathbf{A})$ and on the constants $\tilde{p}_1, ..., \tilde{p}_k$ respectively, such that for all $n \geq n_0(p_1, ..., p_k)$, $a\left(n, L_{(\mathbf{A})}\right) \leq b\left(n, L_{\tilde{p}_1,...,\tilde{p}_k}\right)$.
- the generic constants $\left(L_{p_1,...,p_k}, L_{(\mathbf{A})}, n_0(p_1, ..., p_k), n_0((\mathbf{A}))\right)$ can change from line to another, or even within the same line.

Chapitre 2

The notion of regular contrast

We introduce in this preliminary chapter a new notion in the context of M-estimation, that we call "regular contrast estimation". To be regular on a model M and for a law P, a contrast K must achieve three requirements. First, there exists a unique minimizer of the risk over the model M, which is called the projection of the target to be estimated. Secondly, the contrasted functions of the model, which are the images by the contrast K of the functions of the model, can be expanded into the sum of a constant, a linear part and a quadratic part, when suitably centered by the contrasted projection. Thirdly, the excess risk on M, which is the difference between the risk of functions in M and the risk of the projection of the target, must be close enough - in a certain sense to be defined - to an Hilbertian norm on M, locally around the projection. We give three examples of regular contrast estimation, that will be studied in details along this manuscript. Least-squares estimation on linear models is typically a regular situation. We shall consider the case of least-squares regression, see Chapters 3 and 4, and also least-squares estimation of density, see Chapter 6. But the notion of regular contrast allows to go beyond the case of least-squares estimation. In Chapter 5, we study maximum likelihood estimation of density on histograms. Considering the general case of a M-estimator with regular contrast on an affine or linear model, we derive in Chapter 7 upper and lower bounds in probability *with exact constants* for the true and empirical excess risks of the estimator, at least for a reasonable linear dimension of the model.

Our aim in this manuscript is to tackle the question of the validity of the slope heuristics, that was first formulated by Birgé and Massart [23] in a general Gaussian framework, containing in particular least-squares regression with homoscedastic noise and fixed design. This heuristics claims the existence of a minimal penalty, defined to be the maximum level of penalty under which the model selection procedure totally misbehaves. As soon as a penalty is larger than the minimal one the procedure achieves an oracle inequality. Moreover, the optimal penalty is twice the minimal one, and in this case the leading constant in the oracle inequality is close to one. It happens that the minimal penalty and consequently the optimal penalty can be estimated from the data, due to a jump in the dimension of the selected model occurring around the minimal penalty. Birgé and Massart thus proposed in [23] a data-driven algorithm of calibration of penalties, that leads to an asymptotically optimal procedure in their setting. The slope heuristics and the associated calibration algorithm have then been extended to a more general formulation by Arlot and Massart [10], and proved to be efficient in a least-squares heteroscedastic regression with random design setting, considering linear models of histograms. In Arlot and Massart [10], the quantities of interest are expressed in a general M-estimation setting, and the minimal penalty is identified as the mean of the empirical excess risk over the collection of models. A natural question then arises : Is the slope heuristics valid in general M-estimation and under which general constraints on the models ? In [10], the authors conjecture that the use of histogram models is mainly due to technical issues and that the slope heuristics is valid in least-squares regression for more general models. Recently, Lerasle

[56] also proved the validity of the slope heuristics for least-squares density estimation under rather mild assumptions on the considered linear models. We go a step further in Chapter 4, considering least-squares regression, where we recover results of Arlot and Massart [10] in the histogram case, and extend them to models of piecewise polynomials. We also show that the slope heuristics is valid for maximum likelihood estimation of density on histogram models in Chapter 5. By doing so, we prove that the slope heuristics is valid for another framework than the least-squares one.
The chapter is organized as follows. In Section 2.1 we introduce the problem of M-estimation in a general setting and give a few classical examples. Section 2.1.2 is devoted to the notion of margin conditions, that plays a center role in general M-estimation, and in particular in the Statistical Learning Theory. We then introduce the definition of a regular contrast in Section 2.2 and derive the three examples that we will studied in details along the manuscript. We show in Section 2.2.3 that some margin-like conditions can hold in regular contrast estimation. These conditions are pointed at the projection of the target on the model rather than at the target itself, which can be considered as an advantage when for instance one wants to derive concentration inequalities for the empirical excess risk, as it is the case in a recent paper of Boucheron and Massart [27].

2.1 M-estimation

We state in this section the M-estimation problem in a general setting. In Section 2.1.1 we give some basic definitions and a few classical examples. Section 2.1.2 is devoted to the notion of margin condition that plays a center role in M-estimation and especially in the statistical learning theory. A general introduction to M-estimation can be found in van de Geer [77], see also Massart [61] for a nonasymptotic point of view in the context of model selection.

2.1.1 Definitions and examples

Let $(\mathcal{Z}, \mathcal{T})$ be a measurable space, μ be a probability law on $(\mathcal{Z}, \mathcal{T})$ and let $\xi_1, ..., \xi_n$ be n independent random variables with common law P on $(\mathcal{Z}, \mathcal{T})$. We also consider ξ, a generic random variable of law P, independent of the sample $(\xi_1, ..., \xi_n)$. We denote expectations in a functional way : for a suitable function f

$$Pf = P(f) = \mathbb{E}[f(\xi)]$$
$$\mu f = \mu(f) = \int_{\mathcal{Z}} f d\mu$$

and likewise for

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$$

the empirical distribution associated to the data $(\xi_1, ..., \xi_n)$, we write

$$P_n f = P_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(\xi_i) \ .$$

The variance of $P_n f$ is

$$\mathrm{Var}(f) := \mathbb{V}[f(\xi)] = \mathbb{E}\left[f^2(\xi)\right] - (\mathbb{E}[f(\xi)])^2 \ .$$

The positive part of a real number $x \in \mathbb{R}$ is denoted $(x)_+ := \max\{x \ , \ 0\} \geq 0$ and its negative part is $(x)_- := (-x)_+ = \max\{-x \ , \ 0\} \geq 0$. We naturally extend these definitions to real-valued functions, and for a function f defined on $\mathcal{Z}$ and taking values in $\mathbb{R}$ we write,

$$(f)_+ : z \in \mathcal{Z} \longmapsto (f(z))_+ \quad , \quad (f)_- : z \in \mathcal{Z} \longmapsto (f(z))_- \ .$$

We then denote $L_1^-(P)$ the set of real-valued measurable functions on $(\mathcal{Z},\mathcal{T})$ whose negative part is of finite expectation with respect to P,

$$L_1^-(P) = \left\{ f : \mathcal{Z} \longrightarrow \mathbb{R} \;\; \mathcal{T}\text{-measurable} \; ; \; P(f)_- < +\infty \right\} .$$

Notice that expectation with respect to P is well-defined on $L_1^-(P)$, as we can write for any $f \in L_1^-(P)$,

$$Pf := P(f)_+ - P(f)_- \in \overline{\mathbb{R}} ,$$

where $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$.

Definition 2.1 *A functional K from a set of functions $\mathcal{S}$ to $L_1^-(P)$,*

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow \left\{ f : (\mathcal{Z},\mathcal{T}) \to \mathbb{R}, \; P(f)_- < +\infty \right\} \\ s \longmapsto (Ks : z \longmapsto (Ks)(z)) \end{array} \right. ,$$

*is called a **contrast** if a unique element $s_* \in \mathcal{S}$ exists such that*

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) \quad \text{and} \quad P(Ks_*) < +\infty . \tag{2.1}$$

*The element s_* is the **target**, and for any $s \in \mathcal{S}$, Ks is a **contrasted function** and $P(Ks) \in \overline{\mathbb{R}}$ is the **risk** of s.*

Since we have $Ks_* \in L_1^-(P)$, it is easy to see that the condition $P(Ks_*) < +\infty$ is equivalent to

$$Ks_* \in L_1(P) := \{ f : (\mathcal{Z},\mathcal{T}) \to \mathbb{R}, \; P|f| < +\infty \} .$$

According to (2.1), the target s_* is the minimizer of the risk over the set $\mathcal{S}$. It is an unknown quantity as it depends on the law P of the data and one of the main goals in M-estimation is to estimate the target using the sample $(\xi_1, ..., \xi_n)$. We turn now to the definition of a M-estimator, where M is used for "minimum" - do not confuse with the model M.

Definition 2.2 *Let $(K, \mathcal{S}, P)$ be a triplet such that $K : \mathcal{S} \longrightarrow L_1^-(P)$ is a contrast. Take $M \subset \mathcal{S}$. M is called a **model**. A M-estimator $s_n(M)$, associated to the model M under the contrast K, is defined by*

$$s_n(M) \in \arg\min_{s \in M} P_n(Ks) \quad \text{and} \quad P_n(Ks_n(M)) < +\infty \;\; a.s. \tag{2.2}$$

*where the quantity $P_n(Ks)$ is called the **empirical risk** of s. Since the existence of $s_n(M)$ is not always guaranteed, it is also convenient to define for any $\rho > 0$, the set $\mathcal{V}_n(\rho, M)$ of ρ-almost empirical minimizers in M,*

$$\mathcal{V}_n(\rho, M) := \left\{ s \in M, \; P_n(Ks) \leq \inf_{t \in M} P_n(Kt) + \rho \right\} . \tag{2.3}$$

It is worth noticing that the empirical risk $P_n(Ks)$ is well-defined in (2.2) for any $s \in M$, since as $Ks \in L_1^-(P)$ it holds $-\infty < (Ks)(\xi_i) \leq +\infty$ P-*a.s.* for all $i \in \{1, ..., n\}$. Moreover the condition $P_n(Ks_n(M)) < +\infty$ *a.s.* is equivalent to $Ks_n(M) \in L_1(P_n)$ and ensures that we are not in degenerated the case where for all $s \in M$, $P_n(Ks) = +\infty$.
According to (2.2), a M-estimator $s_n(M)$ estimates the target s_* by minimizing the analog of (2.1) for P_n. More precisely, the unknown law P is replaced by the empirical distribution P_n of the sample $(\xi_1, ..., \xi_n)$ and the set $\mathcal{S}$ is replaced by a convenient subset M. One of the main task in M-estimation is to find a "good" model M, and model selection procedures aim to do so in an accurate and systematical way. The following definition gives a natural measure of performance for a M-estimator, and in particular an "ideal" criterion in model selection context.

Definition 2.3 *Let $(K, \mathcal{S}, P)$ be a triplet such that $K : \mathcal{S} \longrightarrow L_1^-(P)$ is a contrast with target s_*. The **excess risk** $\ell(s_*, s)$ of a function $s \in \mathcal{S}$ is*

$$\ell(s_*, s) := P(Ks) - P(Ks_*) = P(Ks - Ks_*) \geq 0 \ . \tag{2.4}$$

Considering a model $M \subset \mathcal{S}$, and assuming that a M-estimator $s_n(M)$ exists on M, the excess risk of the M-estimator $s_n(M)$, also called the true excess risk, is the random quantity

$$\ell(s_*, s_n(M)) = P(Ks_n(M) - Ks_*) \geq 0 \ . \tag{2.5}$$

We can notice that the excess risk $\ell(s_*, s) \in \overline{\mathbb{R}}_+ := \mathbb{R}_+ \cup \{+\infty\}$ is well-defined for any $s \in \mathcal{S}$ since $Ks \in L_1^-(P)$ and $Ks_* \in L_1(P)$. The smaller is the excess risk of a M-estimator, the better is the M-estimator in terms of excess risk to estimate s_*. Another key quantity is the empirical excess risk of a M-estimator, defined as follows.

Definition 2.4 *Let $(K, \mathcal{S}, P)$ be a triplet such that $K : \mathcal{S} \longrightarrow L_1^-(P)$ is a contrast with target s_*. Considering a model $M \subset \mathcal{S}$, and assuming that a M-estimator $s_n(M)$ exists on M, the **empirical excess risk** of the M-estimator $s_n(M)$ is the random quantity*

$$P_n(Ks_* - Ks_n(M)) \ . \tag{2.6}$$

Notice that the empirical excess risk is well-defined since $P_n(Ks_n(M)) < +\infty$. By definition of a M-estimator $s_n(M)$, see (2.2), the empirical excess risk is nondecreasing with respect to the inclusion on the models $M \subset \mathcal{S}$. The empirical excess risk thus gives a measure of how well a model allows to "fit" the data. This is a nonnegative quantity as soon as, for instance, $s_* \in M$. A central problem in M-estimation, and in particular in model selection of M-estimators via penalization, is to understand the relationship between the empirical excess risk and the true excess risk of a M-estimator, as further explained in Chapters 4 and 7.
Let us now give a few examples of contrasts related to some classical statistical frameworks. We begin with the maybe more classical maximum likelihood estimation of density.

- **Maximum likelihood estimation of density** : assume that P has a density

$$s_* = \frac{dP}{d\mu}$$

with respect to a probability measure μ on $(\mathcal{Z}, \mathcal{T})$ such that $P(\ln s_*)_+ < +\infty$. Then, by taking

$$\mathcal{S} = \left\{ s \geq 0 \ \mathcal{T}\text{-measurable} \ ; \ \int_{\mathcal{Z}} s d\mu = 1 \ \& \ P(\ln s)_+ < +\infty \right\}$$

with the convention $\ln 0 = -\infty$, and the Kullback-Leibler contrast

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1^-(P) \\ s \longmapsto (Ks : z \in \mathcal{Z} \mapsto -\ln(s(z))) \end{array} \right. ,$$

it comes, by Jensen inequality,

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) \ .$$

Moreover, as we always have $P(\ln s)_- < +\infty$, it holds $Ks_* \in L_1(P)$. If we take $M \subset \mathcal{S}$, a M-estimator $s_n(M)$ - if some exists - is the well-known maximum likelihood estimator on M. In this context, for any $s \in \mathcal{S}$, the excess risk

$$\ell(s_*, s) = P(Ks - Ks_*) = \mathcal{K}(s_*, s) := \int_{\mathcal{Z}} s_* \ln\left(\frac{s_*}{s}\right) d\mu$$

is the Kullback-Leibler divergence of the density s with respect to s_*. This example will be further developed in Sections 2.1.2 and 2.2.

- **Least-squares Regression** : assume that $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ for a measurable space $\mathcal{X}$ and that for $\xi = (X, Y)$ of law P it holds

$$Y = s_* (X) + \sigma (X) \varepsilon \ ,$$

with $\mathbb{E}\left[Y^2\right] < +\infty$, $\mathbb{E}\left[\varepsilon \,|X\right] = 0$ and $\mathbb{E}\left[\varepsilon^2 \,|X\right] = 1$. Then $s_* = \mathbb{E}\left[Y \,|X = \cdot\right]$ is the regression function of Y with respect to X and by setting

$$\mathcal{S} = L_2\left(P^X\right) := \left\{s : \mathcal{X} \to \mathbb{R} \ ; \ \mathbb{E}\left[s^2 (X)\right] < +\infty\right\}$$

and defining the least-squares regression contrast to be

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1 (P) \left(\subset L_1^- (P)\right) \\ s \longmapsto \left(Ks : z = (x, y) \in \mathcal{Z} \mapsto (y - s (x))^2\right) \end{array} \right. \ ,$$

it holds

$$s_* = \arg\min_{s \in \mathcal{S}} P (Ks) \ \text{ and } \ \ell (s_*, s) = \|s - s_*\|^2_{L_2(P^X)} := \int_{\mathcal{X}} (s - s_*)^2 (x) \, dP^X \ .$$

So the excess risk is given by the quadratic norm in $L_2\left(P^X\right)$. The M-estimators associated to the least-squares regression contrast are the least-squares estimators. This example will be further developed in Sections 2.1.2 and 2.2.

Typically, the two following examples will fail to be regular in the sense given in Section 2.2, or at least non trivial assumptions would be needed in addition to standard ones.

- **Binary Classification** : assume that $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ for a measurable space $\mathcal{X}$ and let $\xi = (X, Y)$ a random variable of law P. If we set

$$\mathcal{S} = \{s : \mathcal{X} \longrightarrow \{0, 1\} \ \text{ measurable}\} \ ,$$

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1 (P) \left(\subset L_1^- (P)\right) \\ s \longmapsto \left(Ks : z = (x, y) \in \mathcal{Z} \mapsto \mathbf{1}_{\{y \neq s(x)\}}\right) \end{array} \right. \ ,$$

and

$$s_* : x \in \mathcal{X} \longmapsto \mathbf{1}_{\{\mathbb{E}[Y|X=x] \geq 1/2\}} \ ,$$

then the risk $P (Ks) = \mathbb{P} (Y \neq s (X))$ is the probability of misclassification of the "classifier" $s \in \mathcal{S}$ and

$$s_* = \arg\min_{s \in \mathcal{S}} P (Ks) \ .$$

The target s_* is called the Bayes classifier. The problem of optimal model selection, and particularly the theoretical understanding of the slope heuristics remains an open issue in this binary classification setting, see for instance Chapter 7 of [4].

- **Level-set estimation** : let μ be a measure of reference in $(\mathcal{Z}, \mathcal{T})$. Typically, if $\mathcal{Z} \subset \mathbb{R}^n$ then $\mu = \text{Leb}$ is the Lebesgue measure on $\mathcal{Z}$. Let $\mathcal{S} = \mathcal{T}$ and $\lambda > 0$, the goal is to estimate a *level-set* of level λ, also called *generalized λ-cluster* see for example Polonik [65], defined as follows

$$C_\lambda \in \arg\max_{C \in \mathcal{S}} \{P (C) - \lambda \mu (C)\} \ . \tag{2.7}$$

In the case where P admits a density f with respect to μ, we easily see that if

$$\{x \in \mathcal{X}, \ f (x) > \lambda\} \subset C_\lambda \subset \{x \in \mathcal{X}, \ f (x) \geq \lambda\} \ ,$$

then C_λ is a level set of level λ and if in addition $\mu\{x \in \mathcal{X},\ f(x) = \lambda\} = 0$ then the optimization problem (2.7) has an essentially unique solution. By identifying sets and binary functions and take value one if and only if the considered point belongs to the set, level set estimation is an example of M-estimation as stated in Definition 2.1, with contrast

$$K : \begin{cases} \mathcal{S} \longrightarrow L_1(P) \left(\subset L_1^-(P)\right) \\ s \longmapsto (Ks : z \mapsto (Ks)(z) = \lambda\mu(s) - s(z)) \end{cases}$$

which is called the *excess mass* contrast and target $s_*(\cdot) = \mathbf{1}_{\{\cdot \in C_\lambda\}}$ for any C_λ solution of (2.7). Moreover, in the case where P admits a density f with respect to μ, we can write for any $C \in \mathcal{T}$, with the abuse of notations that consists in identifying sets and binary-valued functions,

$$\begin{aligned} \ell(C_\lambda, C) = P(KC) - P(KC_\lambda) &= \int_{C_\lambda} (f - \lambda)\, d\mu - \int_C (f - \lambda)\, d\mu \\ &= \int_{C_\lambda \backslash C} (f - \lambda)\, d\mu - \int_{C \backslash C_\lambda} (f - \lambda)\, d\mu \\ &= \int_{C \Delta C_\lambda} |f - \lambda|\, d\mu\ , \end{aligned}$$

where $A\Delta B := (A\backslash B) \cup (A\backslash B)$ is the symmetric difference between the sets A and B.

2.1.2 Margin conditions

We present now a notion that plays a center role in general analysis of M-estimation, and particularly in problems related to Statistical Learning Theory, as for example binary classification. This notion, that was first introduced by Mammen and Tsybakov [60] in a classification setting, is called *margin condition* and relates the $L_2(P)$ structure of contrasted functions - suitably centered by the contrasted target - to their excess risk, see e.g. [75], [62], [44], [39] and [19] for applications in binary classification and also general bounded contrast minimization settings. Margin conditions allow to derive *fast rates* of convergence for M-estimators, by "localizing" the analysis of such a problem on subsets of interest in a given model, see e.g. [62] and [39] for the use of the related *slicing* or *peeling* techniques and explicit fast rates in terms of various entropy conditions on the model. Depending on the authors, the definition of margin conditions can take sightly different forms, we follow here a definition given in Arlot and Bartlett [9].

Definition 2.5 *Let $(K, \mathcal{S}, P)$ be a triplet such that $K : \mathcal{S} \longrightarrow L_1^-(P)$ is a contrast with target s_* and let $M \subset \mathcal{S}$ be a model. The model M satisfies a* ***margin condition*** *with respect to the law P and for the contrast K if there exists a convex non-decreasing function φ_M on $[0, +\infty)$ with $\varphi_M(0) = 0$, such that for every $s \in M$,*

$$\ell(s_*, s) = P(Ks - Ks_*) \geq \varphi_M\left(\sqrt{\mathrm{Var}(Ks - Ks_*)}\right)\ . \tag{2.8}$$

If there exists a convex non-decreasing function φ on $[0, +\infty)$ with $\varphi(0) = 0$, such that for every $s \in \mathcal{S}$,

$$\ell(s_*, s) = P(Ks - Ks_*) \geq \varphi\left(\sqrt{\mathrm{Var}(Ks - Ks_*)}\right)\ , \tag{2.9}$$

then the contrast K satisfies a margin condition with respect to the law P.

Inequality (2.8) is called "local margin condition" in Arlot and Bartlett [9], whereas Inequality (2.9) is called by the authors "global margin condition". In [9], Arlot and Bartlett address one of the leading problems in classification and learning theory, namely adaptivity to (local) margin

condition *in an effective way*, that they propose to study in a model selection via penalization framework - see [9] for full references on the subject. Adaptation to local margin conditions is called "strong margin adaptivity" in [9], and in most of the literature the considered margin conditions in adaptivity problem are global. The authors show that fully data-dependent procedures such that local Rademacher complexities penalization, can adapt to the local margin when the models are nested. In addition, they show a counter-example when the models are not nested, where no strong adaptation is possible considering *any* penalization method.
We turn now to some examples of margin conditions in the settings defined in Section 2.1.1 above. Classical functions φ or φ_M are of the form κx^β with $\kappa > 0$ and $1 < \beta \leq 2$.

- **Maximum likelihood density estimation** : recall that in this case

$$s_* = \frac{dP}{d\mu} \ , \ (Ks)(\cdot) = -\ln(s(\cdot)) \ \text{ and } \ \ell(s_*, s) = \mathcal{K}(s_*, s) = \int_{\mathcal{Z}} s_* \ln\left(\frac{s_*}{s}\right) d\mu \ .$$

  Take a model $M \subset \mathcal{S}$ such that a positive constant B_M exists satisfying

$$\sup_{s \in M} \left\| \ln\left(\frac{s}{s_*}\right) \right\|_\infty \leq B_M < +\infty \ .$$

  In Lemma 1 of Barron and Sheu [15], it is shown that for any density function s such that $\left\| \ln\left(\frac{s}{s_*}\right) \right\|_\infty$ is finite and any constant c, it holds

$$\mathcal{K}(s_*, s) \geq \frac{1}{2} e^{-\left\| \ln\left(\frac{s}{s_*}\right) \right\|_\infty} \int_{\mathcal{Z}} s_* \left(\ln \frac{s_*}{s}\right)^2 d\mu \tag{2.10}$$

  and

$$\mathcal{K}(s_*, s) \leq \frac{1}{2} e^{\left\| \ln\left(\frac{s}{s_*}\right) - c \right\|_\infty} \int_{\mathcal{Z}} s_* \left(\ln \frac{s_*}{s} - c\right)^2 d\mu \ . \tag{2.11}$$

  We thus have, for any $s \in M$, by using (2.10) and taking $c = P\left(\ln\left(\frac{s_*}{s}\right)\right)$ in (2.11),

$$\frac{e^{2B_M}}{2} \operatorname{Var}(Ks - Ks_*) \geq \ell(s_*, s) \geq \frac{e^{-B_M}}{2} \operatorname{Var}(Ks - Ks_*) \ .$$

  By the second inequality, we thus have a margin condition on M such that for every $x \in [0, +\infty)$, $\varphi_M(x) = e^{-B_M} x^2/2$ is convenient.

- **Least-squares Regression** : recall that in this case, for $\xi = (X, Y)$ of law P it holds

$$Y = s_*(X) + \sigma(X)\varepsilon \ ,$$

  with $\mathbb{E}\left[Y^2\right] < +\infty$, $\mathbb{E}\left[\varepsilon \,|X\right] = 0$ and $\mathbb{E}\left[\varepsilon^2 \,|X\right] = 1$ and for all $s \in \mathcal{S} = L_2\left(P^X\right)$ and all $(x, y) \in \mathcal{Z}$,

$$(Ks)(x, y) = (y - s(x))^2 \ .$$

  Assume that

$$0 \leq \sigma^2(X) \leq A \ \ P^X\text{-a.s.}$$

  and take a model $M \subset \mathcal{S}$ such that a positive constant B_M exist satisfying

$$\sup_{s \in M} \|s - s_*\|_\infty \leq B_M < +\infty \ .$$

Then we can write, for all $s \in M$,

$$\begin{aligned}
\mathrm{Var}\left(Ks - Ks_*\right) &\leq \mathbb{E}\left[\left(\left(Y - s\left(X\right)\right)^2 - \left(Y - s_*\left(X\right)\right)^2\right)^2\right] \\
&= \mathbb{E}\left[\left(2\left(Y - s_*\left(X\right)\right) - \left(s\left(X\right) - s_*\left(X\right)\right)\right)^2 \times \left(s\left(X\right) - s_*\left(X\right)\right)^2\right] \\
&= 4\mathbb{E}\left[\sigma^2\left(X\right)\left(s\left(X\right) - s_*\left(X\right)\right)^2\right] + \mathbb{E}\left[\left(s\left(X\right) - s_*\left(X\right)\right)^4\right] \\
&\leq \left(4A + B_M^2\right)\mathbb{E}\left[\left(s\left(X\right) - s_*\left(X\right)\right)^2\right] \\
&= \left(4A + B_M^2\right)\ell\left(s_*, s\right) \ . \qquad (2.12)
\end{aligned}$$

Hence, the model M satisfies a margin condition with $\varphi_M\left(x\right) = \left(4A + B_M^2\right)^{-1} x^2$ on $[0, +\infty)$.

- **Binary Classification** : we can write in this case, for any $s \in \mathcal{S}$,

$$\begin{aligned}
\ell\left(s_*, s\right) &= \mathbb{E}\left[\mathbf{1}_{\{Y \neq s(X)\}} - \mathbf{1}_{\{Y \neq s_*(X)\}}\right] \\
&= \mathbb{E}\left[\left(2 \times \mathbf{1}_{\{Y = s_*(X)\}} - 1\right)\mathbf{1}_{\{s(X) \neq s_*(X)\}}\right] \\
&= \mathbb{E}\left[\left(2 \times \max\left\{\eta\left(X\right), 1 - \eta\left(X\right)\right\} - 1\right)\mathbf{1}_{\{s(X) \neq s_*(X)\}}\right] \\
&= \mathbb{E}\left[\left|2\eta\left(X\right) - 1\right|\mathbf{1}_{\{s(X) \neq s_*(X)\}}\right] \qquad (2.13)
\end{aligned}$$

where $\eta\left(\cdot\right) = \mathbb{E}\left[Y \,|X = \cdot\right]$ is the regression function of Y with respect to X, and also

$$\begin{aligned}
\mathrm{Var}\left(Ks - Ks_*\right) &\leq \mathbb{E}\left[\left(\mathbf{1}_{\{Y \neq s(X)\}} - \mathbf{1}_{\{Y \neq s_*(X)\}}\right)^2\right] \\
&= \mathbb{E}\left[\left|\mathbf{1}_{\{Y \neq s(X)\}} - \mathbf{1}_{\{Y \neq s_*(X)\}}\right|\right] \\
&= \mathbb{E}\left[\mathbf{1}_{\{s(X) \neq s_*(X)\}}\right] \ . \qquad (2.14)
\end{aligned}$$

From (2.13) and (2.14) we see that the relationship between the variance of the constrasted functions and their excess risk depends on the behavior of the regression function of the label Y with respect to X. More precisely, the more it is bounded away from $1/2$ the stronger the margin will be. If we assume that there exists $\varepsilon_0, c_0 > 0$ and $\alpha \geq 1$ such that

$$\forall \varepsilon \in \left(0, \varepsilon_0\right], \quad \mathbb{P}\left[\left|2\eta\left(X\right) - 1\right| \leq \varepsilon\right] \leq c_0 \varepsilon^{\alpha} \ ,$$

then it is well-known, see Tsybakov [75], that a positive constant $L_{\varepsilon_0, c_0, \alpha}$ exists such that

$$\mathrm{Var}\left(Ks - Ks_*\right) \leq L_{\varepsilon_0, c_0, \alpha}\ell\left(s_*, s\right)^{\frac{\alpha}{1+\alpha}} \ , \qquad (2.15)$$

and so the contrast K achieves a margin condition with $\varphi\left(x\right) = L_{\varepsilon_0, c_0, \alpha}^{-1} x^{\frac{2(1+\alpha)}{\alpha}}$. If $\alpha = +\infty$, that is

$$\mathbb{P}\left[\left|2\eta\left(X\right) - 1\right| \leq h\right] = 0 \qquad (2.16)$$

for some $h > 0$, then it holds

$$h \times \mathrm{Var}\left(Ks - Ks_*\right) \leq \ell\left(s_*, s\right) \ , \qquad (2.17)$$

and so the binary classification contrast satisfies a margin condition with $\varphi\left(x\right) = hx^2$. Inequalities (2.15) and (2.17) allow to interpolate between the pioneering results of Vapnik and Červonenkis based on the VC-dimension, see [84], [83], [85] and [82], where no margin conditions were assumed - see also Lugosi [58] for refinements of these results by chaining techniques - and the "zero-error" case where $h = 1$ in (2.16). In terms of rates of convergence on a fixed model M achieving good enough metric entropy conditions, see

for instance Massart and Nédélec [62], the rate may range from $n^{-1/2}$ when no margin condition is assumed to n^{-1} in the zero-error case, and minimax rates of convergence proportional to $n^{-\beta}$ with $1/2 \leq \beta \leq 1$ can be computed, that depend on the parameters ε_0, C_0,α and h in (2.15) and (2.16) respectively. Moreover, discussions on an extra logarithmic factor in the rates of convergence can be found in [62] and [39]. It can be removed or not depending on the "richness" of the model, see [62], or on the behavior of a local version of Alexander's capacity function, see Section 7.2 of [39].

- **Level set estimation** : The following is a rewriting in terms of margin conditions of assumptions given in Polonik [65]. This formulation may be known by specialists, but we could not find in the literature an explicit "margin condition" in level-set estimation by the excess mass approach. We assume that P admits a density f with respect to the measure of reference μ and that $\mu\{x \in \mathcal{X},\ f(x) = \lambda\} = 0$. Recall that in this case, the target C_λ is unique and for any $C \in \mathcal{S}$,

$$\ell(C_\lambda, C) = \int_{C \Delta C_\lambda} |f - \lambda| \, d\mu \ . \tag{2.18}$$

Moreover we write,

$$\begin{aligned} \mathrm{Var}(KC - KC_\lambda) &= \mathrm{Var}(\lambda\mu(C) - \mathbf{1}_C - \lambda\mu(C_\lambda) + \mathbf{1}_{C_\lambda}) \\ &= \mathrm{Var}(\mathbf{1}_C - \mathbf{1}_{C_\lambda}) \\ &\leq P(C \Delta C_\lambda) = \int_{C \Delta C_\lambda} f d\mu \ . \end{aligned} \tag{2.19}$$

We see from (2.18) and (2.19) that eventual margin conditions depend on the behavior of the density f around the level λ. Assume now that the density f is uniformly bounded on $\mathcal{Z}$, that is a positive constant B exists such that $\|f\|_\infty \leq B$. Moreover assume that there exist $\varepsilon_0, c_0 > 0$ and $\alpha \geq 1$ such that for all $\varepsilon_0 \geq \varepsilon > 0$,

$$\mathbb{P}[|f(\xi) - \lambda| \leq \varepsilon] \leq c_0 \varepsilon^\alpha \ . \tag{2.20}$$

Comparing (2.20) to condition (2.15) in the binary classification setting, we see that f and λ in (2.20) respectively "play the role" of the regression function η and the level $1/2$ in (2.15). This is natural as binary classification reduces to the estimation of the level-set of the regression function of level $1/2$. Now, when $|f(z) - \lambda| > \varepsilon > 0$ for some $z \in \mathcal{Z}$, it holds

$$\frac{B}{\varepsilon} |f(z) - \lambda| > B \geq f(z) ,$$

and so we deduce that for all $\varepsilon_0 \geq \varepsilon > 0$,

$$\begin{aligned} \mathrm{Var}(KC - KC_\lambda) = \int_{C \Delta C_\lambda} f d\mu &\leq \frac{B}{\varepsilon} \int_{C \Delta C_\lambda} |f - \lambda| \, d\mu + P\left(\{|f - \lambda| \leq \varepsilon\} \bigcap C \Delta C_\lambda\right) \\ &\leq \frac{B}{\varepsilon} \ell(C_\lambda, C) + P(|f - \lambda| \leq \varepsilon) \\ &\leq \frac{B}{\varepsilon} \ell(C_\lambda, C) + c_0 \varepsilon^\alpha \qquad \text{by (2.20).} \end{aligned} \tag{2.21}$$

By optimizing upper bound (2.21) with respect to the quantity ε for all $\varepsilon_0 \geq \varepsilon > 0$, we get that there exists a constant $L_{\varepsilon_0, c_0, \alpha} > 0$ such that for all $C \in \mathcal{S}$,

$$\mathrm{Var}(KC - KC_\lambda) \leq L_{\varepsilon_0, c_0, \alpha} \ell(C_\lambda, C)^{\frac{\alpha}{1+\alpha}} \ .$$

In this case the excess mass contrast satisfies a margin condition with respect to the law P with $\varphi(x) = L^{-1}_{\varepsilon_0,c_0,\alpha} x^{\frac{2(1+\alpha)}{\alpha}}$. If $\alpha = +\infty$, that is

$$\mathbb{P}\left[|f(\xi) - \lambda| \leq h\right] = 0$$

for some $h > 0$, then it holds

$$\frac{h}{B} \times \mathrm{Var}\left(KC - KC_\lambda\right) \leq \ell\left(C_\lambda, C\right) \ ,$$

and so the excess mass contrast satisfies a margin condition with $\varphi(x) = \frac{h}{B}x^2$. Using these margin conditions, it is highly likely that existing results in the binary classification setting, such as for instance results given in [62] and [39], could be adapted to level-set estimation. Furthermore, we have introduced here a setting for level-set estimation considering that the level λ is fixed, and a more interesting problem could be to let the level varies. Uniform rates of convergence letting λ vary can be found in [65], see also results and references therein on related problems, and in particular tests of multimodality.

2.2 Regular Contrast Estimation

We have introduced in Section 2.1 the problem of M-estimation in a general setting. We formulate now structural constraints on the contrast, defining a context that we call "regular contrast estimation". We give three examples of regular contrast estimation, namely least-squares regression, least-squares density estimation and maximum likelihood density estimation on histograms. These examples are studied in details in Chapters 3 and 4 for least-squares regression, Chapter 5 is devoted to maximum likelihood density estimation and we address the problem of least-squares density estimation in Chapter 6. The notion of regular contrast allows us to achieve upper and lower bounds in probability with exact constants on the true and empirical excess risks, considering linear or affine models, see Chapter 7.
In Chapter 4, we recover and extend recent results of Arlot and Massart [10] that validate the slope heuristics in least-squares heteroscedastic with random design regression on histograms. Moreover, we also recover some results of Lerasle [56] in least-squares density estimation, see 6. Finally, to our best knowledge theoretical investigations concerning the slope phenomenon were up to now restricted to least-squares frameworks. We go beyond this setting by proving the slope heuristics in maximum likelihood density estimation on histograms in Chapter 5.

2.2.1 Definition of a regular contrast

Let us set

$$L_\infty(P) := \{s : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R} \ \ \mathcal{T}\text{-measurable} \ ; \ \ \|s\|_\infty := \mathrm{essup}_{z\in\mathcal{Z}}\left(|s(z)|\right) < +\infty\}$$

where essup is taken with respect to P, and

$$L_\infty(P) := \left\{s : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R} \ \ \mathcal{T}\text{-measurable} \ ; \ \ \|s\|_2 := \sqrt{P(s^2)} < +\infty\right\} \ .$$

For a subset $A \subseteq \mathbb{R}$, $\mathring{A}$ denotes its interior. The notion of regular contrast is stated as follows.

Definition 2.6 *Let $(K, \mathcal{S}, P)$ be a triplet such that $K : \mathcal{S} \longrightarrow L_1^-(P)$ is a contrast with target s_* and let $M \subset \mathcal{S} \bigcap L_\infty(P)$ be a model. The contrast K is said to be **regular** with respect to the model M and the law P if the following conditions hold. There exists a unique **projection** s_M of s_* on M, defined by*

$$s_M = \arg\min_{s\in M} P(Ks) \quad , \quad P(Ks_M) < +\infty \ . \tag{2.22}$$

For all $s \in M$ and P-almost all $z \in \mathcal{Z}$, the following ***expansion*** *holds,*

$$Ks(z) - Ks_M(z) = \psi_0^s + \psi_{1,M}(z)(s - s_M)(z) + \psi_2(\psi_{3,M}(z)(s - s_M)(z)) \tag{2.23}$$

where ψ_0^s is a constant depending on s but not on z, $\psi_{1,M}$ and $\psi_{3,M}$ are functions defined on $\mathcal{Z}$ not depending on s and not identically equal to 0 satisfying $\psi_{1,M} \in L_2(P)$, $\psi_{3,M} \in L_2(P)$, ψ_2 is a function not depending on s, defined on a subset $\mathcal{D}_2 \subseteq \mathbb{R}$ such that $0 \in \mathring{\mathcal{D}}_2$, $\psi_2(\mathcal{D}_2) \subseteq \overline{\mathbb{R}}$ and $\psi_2(0) = 0$. Moreover, there exist $A_2, L_2 > 0$ such that for all $\delta \in [0, A_2]$, it holds $[-\delta, \delta] \subset \mathcal{D}_2$ and for all $(x, y) \in [-\delta, \delta]$,

$$|\psi_2(x) - \psi_2(y)| \leq L_2 \delta |x - y| \ . \tag{2.24}$$

In addition, we set

$$M_0 = \mathrm{Span}\{s - s_M \ ; \ s \in M\} \ . \tag{2.25}$$

Then there exists an ***Hilbertian norm*** $\|\cdot\|_{H,M}$ *on M_0 and positive constants $A_H, L_H > 0$ such that, for all $t \in M_0$,*

$$\|t\|_2 \leq A_H \|t\|_{H,M} \ . \tag{2.26}$$

Furthermore, for all $\delta \in \left[0, L_H^{-1}\right]$ and all $s \in M$ such that $\|s - s_M\|_\infty \leq \delta \leq L_H^{-1}$, it holds

$$(1 - L_H \delta)\|s - s_M\|_{H,M}^2 \leq P(Ks - Ks_M) \leq (1 + L_H \delta)\|s - s_M\|_{H,M}^2 \ . \tag{2.27}$$

If we can write $\psi_2 \equiv 0$ for all $s \in M$, then the contrast is ***linear*** *and Inequality (2.24) is satisfied, with $\mathcal{D}_2 = \mathbb{R}$ and any $A_2, L_2 > 0$.*

Let us now comment on the previous definition of regular contrast. We ask for three properties. First, we assume that there exists a unique minimizer s_M of the risk on M, given by (2.22), and that the contrasted projection has a finite expectation with respect to P, i.e. $P(Ks_M) < +\infty$. Notice that we always have $P(Ks_M)_- < +\infty$, so $P(Ks_M) > -\infty$. In addition, the property $P(Ks_M) < +\infty$ is equivalent to $Ks_M \in L_1(P)$ and ensures that we are not in the case where for all $s \in M$, $P(Ks_M) = +\infty$. Moreover, this allows to consider the **excess risk on** M, also called the true excess risk on M defined to be, for any $s \in M$,

$$P(Ks - Ks_M) \geq 0 \ .$$

For a M-estimator $s_n(M) \in M$, its **empirical excess risk on** M is

$$P_n(Ks_M - Ks_n(M)) \geq 0 \ .$$

Notice that the excess risk of a M-estimator $s_n(M)$ splits up into the sum of the excess risk of the projection s_M and the excess risk on M of the M-estimator :

$$P(Ks_n(M) - Ks_*) = \underbrace{P(Ks_n(M) - Ks_M)}_{\text{variance term}} + \underbrace{P(Ks_M - Ks_*)}_{\text{bias term}} \ .$$

The excess risk of the projection $P(Ks_M - Ks_*) = \ell(s_*, s_M)$, relates "how far" the model M is from the target s_*. More precisely it quantifies the quality of approximation of the model M with respect to the target s_* in terms of risk, and it is generally called the **bias** of the model M. As we will further see in Chapter 7, the excess risk on M of the M-estimator is related to the "complexity" of the model M, and is essentially a nondecreasing quantity with respect the complexity of the model. It is often called the **variance** term of the excess risk of a M-estimator. Usually, if we take a "large" model suitably chosen according to some prior knowledge, on the regularity of the target s_* for instance, it is likely to have a small bias and a large variance term. On the opposite, a "small" model is on contrary likely to have a smaller variance term but a larger bias. One of the main goal in model selection is to achieve an

accurate trade-off between bias and variance in order to select an estimator with an excess risk as small as possible, see Massart [61].
Our second claim in definition 2.6 is that the contrast K, suitably centered by the contrasted projection Ks_M, can be expanded, see (2.23), into a sum of a constant term, i.e. a term not depending on $z \in \mathcal{Z}$, a linear term and a "quadratic" term, for all function $s \in M$. The condition (2.24) ensures that the term depending on ψ_2 in the expansion of the contrast indeed behaves as a quadratic term. Uniqueness of the expansion is a more technical issue and is discussed in Section 2.2.4 below.
Thirdly, we require that the excess risk on M can be bounded from above and from below by an Hilbertian norm $\|\cdot\|_{H,M}$ as soon as the considered functions $s \in M$ are close enough to the projection s_M in sup-norm. More precisely, the excess risk is equivalent to the Hilbertian norm $\|s - s_M\|_{H,M}$ as s tends to s_M in sup-norm. By stating that $\|\cdot\|_{H,M}$ is an Hibertian norm on M_0 we mean that there exists a inner product $\langle \cdot, \cdot \rangle_{H,M}$ on the vector space M_0 such that for any $t \in M_0$,

$$\langle t, t \rangle_{H,M} = \|t\|_{H,M}^2 \ .$$

Moreover, we ask for a domination of the $L_2(P)$ norm by the Hilbertian norm $\|\cdot\|_{H,M}$ on M_0. In particular, this ensures "local" uniqueness of the projection s_M. Indeed, let $s \in M$ such that $0 < \|s - s_M\|_\infty < L_H^{-1}$, so that it holds

$$P(Ks - Ks_M) \geq (1 - L_H\delta) \|s - s_M\|_{H,M}^2 \geq A_H^{-2} (1 - L_H\delta) \|s - s_M\|_2^2 > 0 \ ,$$

hence $P(Ks) > P(Ks_M) = \inf_{s \in M} P(Ks)$ and the projection s_M is thus the unique minimizer of the risk in the subset

$$M \bigcap \left\{ s \in L_\infty(P) \ ; \ \|s - s_M\|_\infty < L_H^{-1} \right\} .$$

We turn now to the presentation of the three examples of regular contrast estimation that will be developed in this manuscript.

2.2.2 Three examples

Maximum likelihood estimation of density on histograms

Recall that in maximum likelihood estimation of density we have

$$s_* = \frac{dP}{d\mu} \ , \ \mathcal{S} = \left\{ s \geq 0 \ \mathcal{T}\text{-measurable} \ ; \ \int_{\mathcal{Z}} s d\mu = 1 \ \& \ P(\ln s)_+ < +\infty \right\} \ ,$$

and K is the Kullback-Leibler contrast

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1^-(P) \\ s \longmapsto (Ks : z \in \mathcal{Z} \mapsto -\ln(s(z))) \end{array} \right. .$$

We also ask that $Ks_* \in L_1(P)$. Let M be the model of histogram densities associated to a finite partition Λ_M of $\mathcal{Z}$, defined to be

$$M = \left\{ s = \sum_{I \in \Lambda_M} \beta_I \mathbf{1}_I \ ; \ (\beta_I)_{I \in \Lambda_M} \in \mathbb{R}^{D_M}, \ s \geq 0, \ \int_{\mathcal{Z}} s d\mu = 1 \right\}$$

where $D_M = \mathrm{Card}(\Lambda_M)$. We ask that for all $I \in \Lambda_M$, $\mu(I) > 0$. In Chapter 5, where this case is studied in details, M is rather denoted by $\widetilde{M}$. For the histogram model M, the projection s_M exists and is uniquely given by

$$s_M = \sum_{I \in \Lambda_M} \frac{P(I)}{\mu(I)} \mathbf{1}_I \ . \tag{2.28}$$

It is easy to see that $Ks_M \in L_1(P)$. Moreover, the maximum likelihood estimator $s_n(M)$ exists and is P-$a.s.$ unique, satisfying

$$s_n(M) = \sum_{I \in \Lambda_M} \frac{P_n(I)}{\mu(I)} \mathbf{1}_I \ .$$

Notice that if for some $I \in \Lambda_M$, $P_n(I) = 0$ and $P(I) > 0$ then the M-estimator has an infinite risk $P(Ks_n(M)) = +\infty$, or in other words $Ks_n(M) \in L_1^-(P) \setminus L_1(P)$.
Let us introduce $\psi_{1,M}$ and $\psi_{3,M}$ two functions on $\mathcal{Z}$ satisfying

$$\psi_{1,M} = -\psi_{3,M} = -\frac{1}{s_M}$$

and

$$\psi_2 : x \in [-1; +\infty) \, (:= \mathcal{D}_2) \longmapsto \begin{cases} x - \log(1+x) & \text{if} \quad x > -1 \\ +\infty & \text{if} \quad x = -1 \end{cases} .$$

Notice that $0 \in \mathring{\mathcal{D}}_2$, $\psi_2(\mathcal{D}_2) \subseteq \overline{\mathbb{R}}$, $\psi_2(0) = 0$, and if we set $A_2 = 1/2$ then for any $\delta \in [0, A_2]$, it holds $[-\delta, \delta] \subset \mathcal{D}_2$ and for all $(x, y) \in [-\delta, \delta]$,

$$|\psi_2(x) - \psi_2(y)| \leq L_2 \delta |x - y| \ ,$$

with $L_2 = 1$. Moreover as $(\ln s_M) \in L_1(P)$ it holds $s_M > 0$ P-$a.s$, thus for all $s \in M$, we have, with the convention $\ln(0) = -\infty$,

$$\begin{aligned} Ks(z) - Ks_M(z) &= -\ln\left(\frac{s(z)}{s_M(z)}\right) = -\ln\left(1 + \frac{s(z) - s_M(z)}{s_M(z)}\right) \\ &= -\frac{s(z) - s_M(z)}{s_M(z)} + \left(\frac{s(z) - s_M(z)}{s_M(z)} - \ln\left(1 + \frac{s(z) - s_M(z)}{s_M(z)}\right)\right) \\ &= \psi_{1,M}(z)(s - s_M)(z) + \psi_2(\psi_{3,M}(z)(s - s_M)(z)) \qquad P\text{-}a.s. \end{aligned}$$

Hence, the expansion given in (2.23) holds in this case, with $\psi_0^s = 0$ for all $s \in M$. In addition, see Proposition 5.1 of Chapter 5, a Pythagorean-like identity holds for the Kullback-Leibler divergence on the model M, namely

$$\mathcal{K}(s_*, s) = \mathcal{K}(s_*, s_M) + \mathcal{K}(s_M, s) \ \text{, for any } s \in M \ . \tag{2.29}$$

This allows us to show, see Lemma 5.4 of Chapter 5, that if there exists $A_{\min} > 0$ such that $\inf_{z \in \mathcal{Z}} s_*(z) \geq A_{\min} > 0$ then by (2.28) it holds $\inf_{z \in \mathcal{Z}} s_M(z) \geq A_{\min} > 0$ and if we set $L_H = \frac{4}{3A_{\min}} > 0$, then for any $s \in M$ such that $\|s - s_M\|_\infty \leq \delta \leq L_H^{-1}$, it holds

$$(1 - L_H\delta) \frac{1}{2} \left\| \frac{s - s_M}{s_M} \right\|_2^2 \leq P(Ks - Ks_M) \leq (1 + L_H\delta) \frac{1}{2} \left\| \frac{s - s_M}{s_M} \right\|_2^2 .$$

Hence, by setting

$$\|s\|_{H,M} = \frac{1}{\sqrt{2}} \left\| \frac{s}{s_M} \right\|_2 \quad \text{for any } s \in L_2(P) \ ,$$

it follows, since $\inf_{z \in \mathcal{Z}} s_M(z) \geq A_{\min} > 0$, that $\|\cdot\|_{H,M}$ is an Hilbertian norm on $L_2(P)$ and particularly on M_0. Moreover, if $\|s_*\|_\infty < +\infty$ it holds by (2.28) $\|s_M\|_\infty \leq \|s_*\|_\infty < +\infty$ and so for any $s \in L_2(P)$,

$$\|s\|_2 \leq A_H \|s\|_{H,M} \ ,$$

with $A_H = \|s_*\|_\infty / \sqrt{2}$.

From the above calculations, we can conclude that if

$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_M(z) \leq \|s_*\|_\infty < +\infty \ ,$$

then the Kullback-Leibler contrast is regular on the histogram model M under the law P. To extend this result to more general *convex* models than the model of piecewise-constant densities with respect to law μ, we need in particular to generalize the Pythagorean-like identity for the Kullback-Leibler divergence stated in (2.29), see Section 5.4.1 of Chapter 5.

Least-squares regression

In least-squares regression we have $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and for any random variable $\xi = (X, Y)$ of law P,

$$Y = s_*(X) + \sigma(X)\varepsilon \ ,$$

with $\mathbb{E}\left[Y^2\right] < +\infty$, $\mathbb{E}\left[\varepsilon \,|X\right] = 0$ and $\mathbb{E}\left[\varepsilon^2 \,|X\right] = 1$. Moreover, $\mathcal{S} = L_2\left(P^X\right)$, the least-squares regression contrast is

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1(P) \left(\subset L_1^-(P)\right) \\ s \longmapsto \left(Ks : z = (x,y) \in \mathcal{Z} \mapsto (y - s(x))^2\right) \end{array} \right. ,$$

and the excess risk is given by the natural Hilbertian norm in $L_2\left(P^X\right)$,

$$\ell(s_*, s) = \|s - s_*\|^2_{L_2(P^X)} \quad \text{for any } s \in \mathcal{S}.$$

By abuse of notation, we will identify a function s from $\mathcal{X}$ to $\mathbb{R}$ with its extension $\tilde{s}$ to $\mathcal{Z}$, defined as follows

$$\tilde{s} : z = (x, y) \in \mathcal{Z} \longrightarrow \tilde{s}(z) = s(x) \ .$$

This allows in particular to write for any $s \in \mathcal{S}$,

$$\ell(s_*, s) = \|s - s_*\|_2^2 \ .$$

Now, we consider a finite-dimensional vector space M of $L_2\left(P^X\right)$. Clearly there exists a unique orthogonal projection s_M of s_* onto M, and we have the following Pyhagorean identity for any $s \in M$,

$$\|s - s_*\|_2^2 = \|s - s_M\|_2^2 + \|s_M - s_*\|_2^2 \ . \tag{2.30}$$

This allows to deduce that s_M is a projection in the sense of Definition 2.6 for the least-squares contrast, since from (2.30) it comes,

$$\begin{aligned} s_M &= \arg\min_{s \in M} \|s - s_*\|_2^2 \\ &= \arg\min_{s \in M} P(Ks - Ks_*) \\ &= \arg\min_{s \in M} P(Ks) \ . \end{aligned}$$

Moreover, we set for all $z = (x, y) \in \mathcal{Z}$,

$$\psi_{1,M}(z) = -2(y - s_M(x)) \ , \ \psi_{3,M}(z) = 1$$

and

$$\text{for all } u \in \mathbb{R} =: \mathcal{D}_2, \ \psi_2(u) = u^2 \ . \tag{2.31}$$

Hence, from the definition of ψ_2 given in (2.31), we check that

$$0 \in \mathring{\mathcal{D}}_2 \ , \ \psi_2(\mathcal{D}_2) \subseteq \mathbb{R} \ , \ \psi_2(0) = 0 \ ,$$

and for any $A_2 > 0$, by setting $L_2 = 2A_2$, we have for any $\delta \in [0, A_2]$ and for all $(x, y) \in [-\delta, \delta]$,

$$|\psi_2(x) - \psi_2(y)| \leq L_2 \delta |x - y| \ .$$

In addition, the contrast can be expanded as follows, for all $s \in M$ and all $z = (x, y) \in \mathcal{Z}$,

$$\begin{aligned} Ks(z) - Ks_M(z) &= (y - s(x))^2 - (y - s_M(x))^2 \\ &= \psi_{1,M}(z)(s - s_M)(z) + \psi_2(\psi_{3,M}(z)(s - s_M)(z)) \ . \end{aligned}$$

Hence, the expansion given in (2.23) holds in this case, with $\psi_0^s = 0$ for all $s \in M$. Moreover, by the Pythagorean identity (2.30) we have

$$P(Ks - Ks_M) = \|s - s_M\|_2^2 \ ,$$

and thus, by setting $\|\cdot\|_{H,M} = \|\cdot\|_2$, inequality (2.27) is satisfied for any $L_H > 0$, $A_H \geq 1$, so we can conclude that the least-squares regression contrast is regular for the model M and law P, with $A_H = 1$ and any $A_2, L_H > 0$. This example is studied in details in Chapters 3 and 4.

Least-squares estimation of density

Least-squares estimation of density is defined as follows. Let μ be a known probability measure on $(\mathcal{Z}, \mathcal{T})$ and assume that P admits a density f with respect to μ :

$$f = \frac{dP}{d\mu} \ .$$

Let endow the space of square integrable measurable functions with respect to the law μ, namely

$$L_2(\mu) = \left\{ s \ , \ \mu\left(s^2\right) < +\infty \right\} \ ,$$

with its natural Hilbertian structure associated to the inner product

$$\langle s, t \rangle = \mu(st) = \int_{\mathcal{Z}} st d\mu$$

and the Hilbertian norm $\|.\|$ defined by

$$\|s\|^2 = \|s\|_{L_2(\mu)}^2 = \langle s, s \rangle = \mu\left(s^2\right) = \int_{\mathcal{Z}} s^2 d\mu \ .$$

Moreover, we assume that there exists a function s_0, typically $s_0 \equiv 1$ if $\mathcal{Z}$ is the unit interval or $s_0 \equiv 0$, and another function s_* such that

$$f = s_0 + s_* \ \text{ and } \int_{\mathcal{Z}} s_* s_0 d\mu = 0 \ .$$

We also define the orthogonal vector space of s_0 in $L_2(\mu)$,

$$\{s_0\}^\perp = \{ s \in L_2(\mu) \ , \ \langle s, s_0 \rangle = 0 \} \ .$$

Thus we have $s_* \in \{s_0\}^\perp$, and s_* is the target. Now, let $s \in \{s_0\}^\perp$, we have

$$\begin{aligned} \|s - s_*\|^2 &= \|s\|^2 - 2\langle s, s_* \rangle + \|s_*\|^2 \\ &= \|s\|^2 - 2\langle s, f \rangle + \|s_*\|^2 \\ &= \|s\|^2 - 2Ps + \|s_*\|^2 \end{aligned}$$

and we deduce that

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks)$$

where $\mathcal{S} := \{s_0\}^\perp$ and the least-squares density contrast $K : L_2(\mu) \longrightarrow L_1(P)$ satisfies

$$Ks = \|s\|^2 - 2s \ , \text{ for all } s \in L_2(\mu) .$$

Now, let us take a finite dimensional vector space $M \subset \mathcal{S}$. For every $s \in M$,

$$\langle s, s_0 \rangle = \int_{\mathcal{Z}} s s_0 d\mu = 0 \ .$$

The considered estimator on M is the least-squares estimator, defined as follows

$$\begin{aligned} s_n(M) &\in \arg\min_{s \in M} P_n(Ks) \\ &= \arg\min_{s \in M} \left\{ \|s\|^2 - 2P_n s \right\} \ . \end{aligned}$$

It is easy to check that such an estimator exists and is unique, and if D is the linear dimension of $(M, \|\cdot\|)$ and $(\varphi_k)_{k=1}^D$ is an orthonormal basis of $(M, \|\cdot\|)$, then

$$s_n(M) = \sum_{k=1}^{D} P_n(\varphi_k) \varphi_k \ .$$

Moreover, notice that for any $s \in \{s_0\}^\perp$,

$$\begin{aligned} P(Ks - Ks_*) &= PKs - PKs_* \\ &= \|s\|^2 - 2\langle s, f \rangle - \|s_*\|^2 + 2\langle s_*, f \rangle \\ &= \|s\|^2 - 2\langle s, s_* \rangle + \|s_*\|^2 \\ &= \|s - s_*\|^2 \geq 0 \ , \end{aligned}$$

and so the excess risk $P(Ks - Ks_*)$ is the $L_2(\mu)$ loss. If we denote by s_M the linear projection of s_* onto M in $L_2(\mu)$, we then have

$$PKs_M - PKs_* = \inf_{s \in M} \{PKs - PKs_*\} \ , \tag{2.32}$$

and from (2.32), we deduce that s_M is the unique projection of s_* onto M in the sense of Definition 2.6,

$$s_M = \arg\min_{s \in M} PK(s) \ .$$

We also notice that by the Pythagorean theorem we have for all $s \in M$,

$$\|s - s_*\|^2 = \|s - s_M\|^2 + \|s_M - s_*\|^2 \ ,$$

and so it holds for all $s \in M$,

$$P(Ks - Ks_M) = \|s - s_M\|^2 \geq 0 \ .$$

We thus set $\|\cdot\|_{H,M} = \|\cdot\|$, and we easily see that Inequality (2.27) of Definition 2.6 is satisfied for any $L_H > 0$. If we assume that $\|f\|_\infty < +\infty$, then it moreover holds for any $s \in M$,

$$\|s\|_2 \leq A_H \|s\|_{H,M}$$

with $A_H = \|f\|_\infty$. Finally, by setting

$$\psi_{1,M} \equiv -2$$
$$\psi_0^s = \|s\|^2 - \|s_M\|^2$$

we can write for any $s \in M$, and any $z \in \mathcal{Z}$,

$$Ks(z) - Ks_M(z) = \psi_0^s + \psi_{1,M}(z)(s - s_M)(z) \ .$$

From the computations above, we conclude that if $\|f\|_\infty < +\infty$, then the least-squares density contrast is linear for the model M and law P. Least-squares density estimation is studied in Chapter 6, where we also consider the case where the density f is only assumed to belong to $L_2(\mu)$.

2.2.3 Margin-like conditions in regular contrast estimation

We intend to show in Proposition 2.1 below that margin-like conditions can hold in regular contrast estimation. The margin-like conditions are pointed at the projection s_M of the target s_* onto a model M rather than at the target s_* itself, as it is usually the case when considering standard margin conditions, see Defintion 2.5 above.

Proposition 2.1 *Let $(K, \mathcal{S}, P)$ be a triplet such that $K : \mathcal{S} \longrightarrow L_1^-(P)$ is a contrast with target s_* and let $M \subset \mathcal{S} \bigcap L_\infty(P)$ be a model such that K is a regular contrast with respect to the model M and the law P. Assume that $\|\psi_{1,M}\|_\infty < +\infty$ and $\|\psi_{3,M}\|_\infty < +\infty$, where for any $s \in M$ and P-almost all $z \in \mathcal{Z}$,*

$$Ks(z) - Ks_M(z) = \psi_0^s + \psi_{1,M}(z)(s - s_M)(z) + \psi_2(\psi_{3,M}(z)(s - s_M)(z)) \ .$$

Then there exist two positive constants $A_M, B_M > 0$ such that for any $s \in M$ satisfying $\|s - s_M\|_\infty \leq A_M$, it holds

$$\mathrm{Var}(Ks - Ks_M) \leq B_M \times P(Ks - Ks_M) \ . \tag{2.33}$$

Moreover, with notations of Definition 2.6, (2.33) hold with

$$A_M = \left(L_H^{-1}/2\right) \wedge \left(A_2 / \|\psi_{3,M}\|_\infty^2\right) > 0$$

and

$$B_M = 4\left(\|\psi_{1,M}\|_\infty^2 + \left(L_2 A_2 \|\psi_{3,M}\|_\infty\right)^2\right) A_H^2 > 0 \ .$$

Let us comment on Proposition 2.1. We emphasize on the fact that relations of the form of (2.33) are convenient when for instance, one wants to prove concentration inequalities for the empirical excess risk $P_n(Ks_M - Ks_n(M))$, as studied in Boucheron and Massart [27]. Results of [27] have also been applied in Arlot and Massart [10] in the case of least-squares heteroscedastic bounded regression with random design on finite-dimensional models of histograms. For instance, in Proposition 10 of [10] transposed in our notations, the authors show that if $|Y| \leq A < +\infty$ *a.s.* then for every $x \geq 0$ there exists an event of probability at least $1 - e^{1-x}$ on which for every $\theta \in (0,1)$,

$$\begin{aligned} &|P_n(Ks_M - Ks_n(M)) - \mathbb{E}[P_n(Ks_M - Ks_n(M))]| \\ &\leq L\left[\theta \ell(s_*, s_M) + \frac{A^2\sqrt{D_M}\sqrt{x}}{n} + \frac{A^2 x}{\theta n}\right] \end{aligned} \tag{2.34}$$

for some absolute constant L, where D_M denotes the linear dimension of the model of histograms M. The bias term $\ell(s_*, s_M)$ seems rather unnatural in (2.34) as one could expect a better concentration of the empirical excess risk $P_n(Ks_M - Ks_n(M))$ for small models, which on the contrary have a large bias. In the proof of Proposition 10 given in [10], the bias term arises from the use of the margin condition that holds in this case and that is given in (2.12) above. Inequality (2.12) can in fact be replaced in this case by an inequality of the form of (2.33), and so the bias term in (2.34) could be removed, with only a change in numerical constants for the other terms in (2.34), by a straightforward adaptation of the proof given in [10]. Least-squares heteroscedastic regression with random design on a finite-dimensional model is indeed an example of regular contrast estimation as shown in Section 2.2.2. Moreover, with notations of Defintion 2.6, we can take in this case $A_H = 1$ and any $A_2, L_H > 0$. In the particular case of an histogram model on a finite partition Λ_M with $D_M := \mathrm{Card}(\Lambda_M)$,

$$M = \left\{ s = \sum_{I \in \Lambda_M} \beta_I \mathbf{1}_I \ ; \ (\beta_I)_{I \in \Lambda_M} \in \mathbb{R}^{D_M} \right\} ,$$

it is easy to see that the projection s_M of the regression function s_* can be written

$$s_M(x) = \sum_{I \in \Lambda_M} \frac{\mathbb{E}[Y \mathbf{1}_{X \in I}]}{P^X(I)} \mathbf{1}_{x \in I} \ , \text{ for any } x \in \mathcal{X} \tag{2.35}$$

and that the least-squares estimator is given by

$$s_n(M) = \sum_{I \in \Lambda_M} \frac{P_n(Y \mathbf{1}_{X \in I})}{P_n(\mathbf{1}_{X \in I})} \mathbf{1}_I \ . \tag{2.36}$$

If we assume that there exists $A > 0$ such that

$$|Y| \leq A < +\infty \ a.s. \tag{2.37}$$

then

$$\|s_M\|_\infty \leq A \text{ and } \|s_n(M)\|_\infty \leq A \ a.s. \tag{2.38}$$

and we can indeed restrict the analysis of the problem in the subset $\{s \in M \ ; \ \|s - s_M\|_\infty \leq 2A\}$. Moreover, from (2.37) and (2.38), since in the regression case $\psi_{1,M}(x,y) = -2(y - s_M(x))$ for any $(x,y) \in \mathcal{Z}$, it follows that

$$\|\psi_{1,M}\|_\infty \leq 4A \ ,$$

and we can easily see from Proposition 2.1 that for any $s \in M$ such that $\|s - s_M\|_\infty \leq 2A$, it holds

$$\mathrm{Var}(Ks - Ks_M) \leq 12A^2 P(Ks - Ks_M) \ .$$

In the same spirit, each time that the "Noise Condition" stated in Boucheron and Massart [27] can be pointed at the projection of the target rather than at the target itself this could allow to remove the bias of the considered model from the derived concentrations inequalities for the empirical excess risk.

To conclude, some margin-like conditions hold in our regular framework under the assumption that the function $\psi_{1,M}$ is uniformly bounded on $\mathcal{Z}$, but it seems that this not sufficient to derive upper and lower bounds for the excess risks on a fixed model, *with exact constants*. As a matter of fact, in Chapter 7, where we address such a problem in a regular contrast estimation setting, we avoid the use of margin-like inequalities stated in Proposition 2.1, by taking advantage of the linearity of the considered model.

We give now the proof of Proposition 2.1.

Proof. From Definition 2.6, there exist an Hilbertian norm $\|\cdot\|_{H,M}$ on M_0 and positive constants $A_2, L_2, A_H, L_H > 0$ such that for all $\delta \in [0, A_2]$ and all $(x, y) \in [-\delta, \delta]$,

$$|\psi_2(x) - \psi_2(y)| \leq L_2 \delta |x - y| \ , \tag{2.39}$$

$$\|\cdot\|_2 \leq A_H \|\cdot\|_{H,M} \tag{2.40}$$

and for all $s \in M$ such that $\|s - s_M\|_\infty \leq \delta \leq L_H^{-1}$, it holds

$$(1 - L_H \delta) \|s - s_M\|_{H,M}^2 \leq P(Ks - Ks_M) \ . \tag{2.41}$$

Moreover $0 < \|\psi_{3,M}\|_\infty < +\infty$ since $\psi_{3,M} \in L_\infty(P)$ is non identically equal to zero, and we take

$$A_M = \left(L_H^{-1}/2\right) \wedge \left(A_2 / \|\psi_{3,M}\|_\infty\right) > 0.$$

Hence, for any $s \in M$ such that $\|s - s_M\|_\infty \leq A_M$, for any $z \in \mathcal{Z}$, by (2.39) applied with $x = \psi_{3,M}(z)(s - s_M)(z)$, $y = 0$ and $\delta = A_M \|\psi_{3,M}\|_\infty$ it holds, since $\psi_2(y) = \psi_2(0) = 0$,

$$\begin{aligned} |\psi_2(\psi_{3,M}(z)(s - s_M)(z))| &\leq L_2 A_2 |\psi_{3,M}(z)(s - s_M)(z)| \\ &\leq L_2 A_2 \|\psi_{3,M}\|_\infty |(s - s_M)(z)| \ . \end{aligned} \tag{2.42}$$

In addition, by (2.41) applied with $A_M \leq L_H^{-1}/2 < L_H^{-1}$ and by (2.40), it holds for any $s \in M$ such that $\|s - s_M\|_\infty \leq A_M$,

$$\|s - s_M\|_2^2 \leq A_H^2 \|s - s_M\|_{H,M}^2 \leq 2A_H^2 P(Ks - Ks_M) \ . \tag{2.43}$$

Now, for any $s \in M$ such that $\|s - s_M\|_\infty \leq A_M$,

$$\begin{aligned} \mathrm{Var}(Ks - Ks_M) &= \mathrm{Var}(\psi_0^s + \psi_{1,M} \cdot (s - s_M) + \psi_2 \circ (\psi_{3,M} \cdot (s - s_M))) \\ &= \mathrm{Var}(\psi_{1,M} \cdot (s - s_M) + \psi_2 \circ (\psi_{3,M} \cdot (s - s_M))) \\ &\leq P(\psi_{1,M} \cdot (s - s_M) + \psi_2 \circ (\psi_{3,M} \cdot (s - s_M)))^2 \\ &\leq 2P(\psi_{1,M} \cdot (s - s_M))^2 + 2P(\psi_2 \circ (\psi_{3,M} \cdot (s - s_M)))^2 \\ &\leq 2\|\psi_{1,M}\|_\infty^2 \|s - s_M\|_2^2 + 2\left(L_2 A_2 \|\psi_{3,M}\|_\infty\right)^2 \|s - s_M\|_2^2 \qquad \text{by (2.42)} \\ &\leq 4\left(\|\psi_{1,M}\|_\infty^2 + \left(L_2 A_2 \|\psi_{3,M}\|_\infty\right)^2\right) A_H^2 P(Ks - Ks_M) \ . \end{aligned}$$

So $B_M = 4\left(\|\psi_{1,M}\|_\infty^2 + \left(L_2 A_2 \|\psi_{3,M}\|_\infty\right)^2\right) A_H^2 > 0$ gives the result. ■

2.2.4 On the uniqueness of the expansion of a regular contrast

In this section, we consider a regular contrast K for a model M and a law P, and we discuss the uniqueness of the parameters appearing in the expansion of the contrast, namely ψ_0^s, $\psi_{1,M}$, $\psi_{3,M}$ and ψ_2 with the notations of Definition 2.6.

Framework :

We assume that for all $z \in \mathcal{Z}$,

$$Ks(z) - Ks_M(z) = \psi_0^s + \psi_{1,M}(z)(s - s_M)(z) + \psi_2(\psi_{3,M}(z)(s - s_M)(z)) \tag{2.44}$$

and

$$Ks(z) - Ks_M(z) = \widetilde{\psi}_0^s + \widetilde{\psi}_{1,M}(z)(s - s_M)(z) + \widetilde{\psi}_2\left(\widetilde{\psi}_{3,M}(z)(s - s_M)(z)\right) \ , \tag{2.45}$$

where ψ_0^s, $\widetilde{\psi}_0^s$ are constants depending on s but not on z, $\psi_{1,M}$, $\widetilde{\psi}_{1,M}$, $\psi_{3,M}$ and $\widetilde{\psi}_{3,M}$ are functions defined on $\mathcal{Z}$ not depending on s and not identically equal to 0 satisfying $\psi_{1,M}$, $\widetilde{\psi}_{1,M} \in L_2(P)$ and $\psi_{3,M}$, $\widetilde{\psi}_{3,M} \in L_\infty(P)$. Moreover ψ_2 and $\widetilde{\psi}_2$ are functions not depending on s, respectively defined on $\mathcal{D}_2 \subseteq \mathbb{R}$ and $\widetilde{\mathcal{D}}_2 \subseteq \mathbb{R}$. We also have $0 \in \mathring{\mathcal{D}}_2 \bigcap \overset{\circ}{\widetilde{\mathcal{D}}}_2$, $\psi_2(\mathcal{D}_2) \bigcup \widetilde{\psi}_2\left(\widetilde{\mathcal{D}}_2\right) \subseteq \overline{\mathbb{R}}$ and $\psi_2(0) = \widetilde{\psi}_2(0) = 0$. Moreover, there exist $A_2, \tilde{A}_2, L_2, \tilde{L}_2 > 0$ such that for all $\delta \in [0, A_2]$, it holds $[-\delta, \delta] \subset \mathcal{D}_2$ and for all $(x, y) \in [-\delta, \delta]$,

$$|\psi_2(x) - \psi_2(y)| \leq L_2 \delta |x - y| \ , \tag{2.46}$$

and also, for all $\delta \in \left[0, \tilde{A}_2\right]$, we have $[-\delta, \delta] \subset \widetilde{\mathcal{D}}_2$ and for all $(x, y) \in [-\delta, \delta]$,

$$\left|\widetilde{\psi}_{22}(x) - \widetilde{\psi}_{22}(y)\right| \leq \tilde{L}_2 \delta |x - y| \ . \tag{2.47}$$

Furthermore, we need the three following definitions.

Definition 2.7 *The **support** $\mathcal{Z}_M$ of the model M on $\mathcal{Z}$ is defined to be*

$$\mathcal{Z}_M := \{z \in \mathcal{Z} \ ; \ \exists s \in M, \ s(z) - s_M(z) \neq 0\} \ . \tag{2.48}$$

Definition 2.8 *There exists a unique partition $\mathcal{P}_M = \left(\mathcal{P}_M^i\right)_{i \in I}$ of $\mathcal{Z}$, called the **discriminative partition** of $\mathcal{Z}$ with respect to the model M, such that*

$$\forall i \in I, \ \forall (z_1, z_2) \in \mathcal{P}_M^i \times \mathcal{P}_M^i, \ \forall s \in M, \ (s - s_M)(z_1) = (s - s_M)(z_2) \ .$$

and

$$\forall (i, j) \in I \times I, \ i \neq j, \ \exists (z_i, z_j) \in \mathcal{P}_M^i \times \mathcal{P}_M^j, \ \exists s \in M, \ (s - s_M)(z_i) \neq (s - s_M)(z_j) \ .$$

If the discriminative partition $\mathcal{P}_M$ of $\mathcal{Z}$ satisfies the following property,

$$\forall (i, j) \in I \times I, \ i \neq j, \ \forall (z_i, z_j) \in \mathcal{P}_M^i \times \mathcal{P}_M^j, \ \forall (a, b) \in \mathbb{R}_* \times \mathbb{R}, \ \exists s \in M, \\ a(s - s_M)(z_i) \neq b(s - s_M)(z_j) \ ,$$

*then it is said to be **totally discriminative**.*

Existence and uniqueness of the discriminative partition of $\mathcal{Z}$ with respect to the model M are straightforward and left to the reader. Notice that $(\mathcal{Z}_M)^c = \mathcal{Z} \backslash \mathcal{Z}_M$ is an element of the discriminative partition. The set I which index the partition is not necessarily finite and we can only say that I is of cardinality not larger than the cardinality of $\mathcal{Z}$. If for instance, M is a model of histograms on a partition Λ_M of $\mathcal{Z}$, then $\Lambda_M = \mathcal{P}_M$ and this discriminative partition is totally discriminative.

Definition 2.9 *The model M is **star-shaped** at the point $s_e \in M$ if for any $s \in M$ and any $\lambda \in [0, 1]$,*

$$t_{s,\lambda} := \lambda(s - s_e) + s_e \in M \ .$$

In the following, we consider a model M which is star-shaped at the projection s_M of the target onto the model. Notice that linear models M are star-shaped at any point $s_e \in M$ in the sense of Definition 2.9. Affine models, that is models M satisfying the fact that there exists $s_a \in M$ such that $\{s - s_a \ ; \ s \in M\}$ is a linear vector space, are also star-shaped at any point $s_e \in M$. More general non-linear models such as linear combinations of sigmoidal functions used in neural networks, see Barron [14], or histograms generated by any partition on $[0, 1]$ into D subintervals, are star-shaped at zero.

Unicity of the constant terms :

If $\mathcal{Z}_M \subsetneq \mathcal{Z}$, then for any $z_c \in \mathcal{Z} \backslash \mathcal{Z}_M$, for any $s \in M$, it holds

$$Ks\left(z_c\right) - Ks_M\left(z_c\right) = \psi_0^s \ , \tag{2.49}$$

since $\psi_2\left(0\right) = 0$, and $\psi_{1,M}\left(z_c\right)$, $\psi_{3,M}\left(z_c\right)$ can take any value. We emphasize on the fact that for any $z \in \mathcal{Z}$, $Ks\left(z\right) = \left(Ks\right)\left(z\right)$ is not necessarily a function of $s\left(z\right)$ and so the constant term ψ_0^s in (2.44) is not necessarily equal to zero. For instance, in least-squares density estimation $\left(Ks\right)\left(z\right)$ is not in general a function of $s\left(z\right)$ when K is the least-squares density contrast, as we have for all $s \in M$, $\psi_0^s = \left\|s\right\|^2 - \left\|s_M\right\|^2$, see Section 2.2.2. However, if $\mathcal{Z} \backslash \mathcal{Z}_M \neq \varnothing$, then ψ_0^s is uniquely defined for all $s \in M$, by formula (2.49). If $\mathcal{Z}_M = \mathcal{Z}$, then the constants ψ_0^s are not necessarily uniquely defined, apart for functions $s \in M$ that vanish at one point at least. More precisely, if there exists $z_0 \in \mathcal{Z}$ such that

$$\left(s - s_M\right)\left(z_0\right) = 0$$

then we have

$$Ks\left(z_0\right) - Ks_M\left(z_0\right) = \psi_0^s$$

and so ψ_0^s is uniquely defined. Moreover, in this case, for any $\lambda \in \left(0, 1\right]$,

$$\left(t_{s,\lambda} - s_M\right)\left(z_0\right) = \lambda\left(s - s_M\right)\left(z_0\right) = 0$$

and so $\psi_0^{t_{s,\lambda}}$ is again uniquely defined, using (2.49).

Unicity of the function $\psi_{1,M}$:

Proposition 2.2 *Assume that the model M is star-shaped at s_M and let $\mathcal{P}_M = \left(\mathcal{P}_M^i\right)_{i \in I}$ be the discriminative partition of $\mathcal{Z}$ with respect to the model M. The function $\psi_{1,M} - \widetilde{\psi}_{1,M}$ is constant on the elements $\mathcal{P}_M^i$ of $\mathcal{P}_M$ such that $\mathcal{P}_M^i \subset \mathcal{Z}_M$. Moreover, for any $s \in M$, there exists a constant $a \in \mathbb{R}$ such that for all $z \in \mathcal{Z}$,*

$$\widetilde{\psi}_{1,M}\left(z\right)\left(s - s_M\right)\left(z\right) = \psi_{1,M}\left(z\right)\left(s - s_M\right)\left(z\right) + a \ . \tag{2.50}$$

if $\mathcal{Z} \backslash \mathcal{Z}_M \neq \varnothing$, then $\psi_{1,M}$ and $\widetilde{\psi}_{1,M}$ are identically equal on $\mathcal{Z}_M$. Otherwise, if $\mathcal{P}_M$ is totally discriminative, then $\psi_{1,M}$ and $\widetilde{\psi}_{1,M}$ are identically equal on $\mathcal{Z} = \mathcal{Z}_M$ or differ on only one element of $\mathcal{P}_M$.

Proposition 2.2 ensures that, when the model M is star-shaped at the projection s_M, the function $\psi_{1,M}$ defined in the expansion of the regular contrast K is uniquely defined on $\mathcal{Z}_M$, at least up to a constant over one element of $\mathcal{P}_M$, when $\mathcal{P}_M$ is assumed to be totally discriminative. In Chapter 7, we will see that the rates of convergence to zero of the excess risks on a fixed model depend on a quantity, called the complexity of the model M, that relates the structure of the model with the function $\psi_{1,M}$ under the law P. Identity (2.50) of Proposition 2.2 allows to show that the complexity is indeed independent of the choice of the function $\psi_{1,M}$, see Remark 7.2 of Section 7.2.2 in Chapter 7.

Remark 2.1 *If $\mathcal{Z} = \mathcal{Z}_M$ and if there exists $z \in \mathcal{Z}$ such that $\psi_{1,M}\left(z\right) \neq \widetilde{\psi}_{1,M}\left(z\right)$, then for any $s \in M$, $\widetilde{\psi}_0^s$ can take the value $\psi_0^s + \left(\psi_{1,M} - \widetilde{\psi}_{1,M}\right)\left(z\right)\left(s - s_M\right)\left(z\right)$ without any contradiction, since the function $\left(\psi_{1,M} - \widetilde{\psi}_{1,M}\right)\left(z\right)\left(s - s_M\right)\left(z\right)$ is constant when z varies in $\mathcal{Z}$, by (2.50). In other words, if $\mathcal{Z} = \mathcal{Z}_M$ and if $s \in M$ is such that for any $z \in \mathcal{Z}$, $\left(s - s_M\right)\left(z\right) \neq 0$, then ψ_0^s can take any value.*

We turn now to the proof of Proposition 2.2. The following lemma will be convenient.

Lemma 2.1 *For any $z_0 \in \mathcal{Z}$ and $s_0 \in M$,*

$$\psi_2\left(\psi_{3,M}\left(z_0\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right)\right) \ll \lambda \text{ as } \lambda \to 0 \ . \tag{2.51}$$

and

$$\widetilde{\psi}_2\left(\widetilde{\psi}_{3,M}\left(z_0\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right)\right) \ll \lambda \text{ as } \lambda \to 0 \ . \tag{2.52}$$

Moreover, if

$$\left(\widetilde{\psi}_{1,M}\left(z_0\right) - \psi_{1,M}\left(z_0\right)\right)\left(s_0 - s_M\right)\left(z_0\right) \neq 0$$

then

$$\psi_0^{t_{s_0,\lambda}} - \widetilde{\psi}_0^{t_{s_0,\lambda}} \sim \lambda\left(\widetilde{\psi}_{1,M}\left(z_0\right) - \psi_{1,M}\left(z_0\right)\right)\left(s_0 - s_M\right)\left(z_0\right) \text{ as } \lambda \to 0 \ . \tag{2.53}$$

Proof of Lemma 2.1. For any $\lambda \in (0,1]$, we deduce from (2.44) and (2.45) applied at z_0 that

$$\begin{aligned}\psi_0^{t_{s_0,\lambda}} - \widetilde{\psi}_0^{t_{s_0,\lambda}} = &\left(\widetilde{\psi}_{1,M}\left(z_0\right) - \psi_{1,M}\left(z_0\right)\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right) + \widetilde{\psi}_2\left(\widetilde{\psi}_{3,M}\left(z_0\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right)\right) \\ &-\psi_2\left(\psi_{3,M}\left(z_0\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right)\right) \ .\end{aligned} \tag{2.54}$$

Moreover,

$$\left(\widetilde{\psi}_{1,M}\left(z_0\right) - \psi_{1,M}\left(z_0\right)\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right) = \lambda\left(\widetilde{\psi}_{1,M}\left(z_0\right) - \psi_{1,M}\left(z_0\right)\right)\left(s_0 - s_M\right)\left(z_0\right) \neq 0 \ . \tag{2.55}$$

In addition, for $\lambda \geq 0$ small enough, it holds for any $s \in M$,

$$\left|\psi_{3,M}\left(z_0\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right)\right| = \lambda\left|\psi_{3,M}\left(z_0\right)\left(s_0 - s_M\right)\left(z_0\right)\right| \in [0, A_2] \tag{2.56}$$

Since $\psi_2\left(0\right) = 0$, we get for λ small enough, by (2.56) and (2.46) applied with $\delta = \lambda\left|\psi_{3,M}\left(z_0\right)\left(s_0 - s_M\right)\left(z_0\right)\right|$, $x = \psi_{3,M}\left(z_0\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right)$ and $y = 0$,

$$\begin{aligned}\left|\psi_2\left(\psi_{3,M}\left(z_0\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right)\right)\right| &\leq L_2\lambda\left|\psi_{3,M}\left(z_0\right)\left(s_0 - s_M\right)\left(z_0\right)\right| \times \left|\psi_{3,M}\left(z_0\right)\left(t_{s_0,\lambda} - s_M\right)\left(z_0\right)\right| \\ &\leq L_2\lambda^2\left|\psi_{3,M}\left(z_0\right)\left(s_0 - s_M\right)\left(z_0\right)\right|^2 \ .\end{aligned} \tag{2.57}$$

We deduce from (2.57) that (2.51) hold and by the same type of arguments (2.52) also hold. Hence, by (2.54), (2.55), (2.51) and (2.52) we deduce that (2.53) hold. ■

Proof of Proposition 2.2. Let us first prove that the function $\psi_{1,M} - \widetilde{\psi}_{1,M}$ is constant on the elements $\mathcal{P}_M^i$ of $\mathcal{P}_M$ such that $\mathcal{P}_M^i \subset \mathcal{Z}_M$. Let $i \in I$ be fixed. If $\psi_{1,M}$ and $\widetilde{\psi}_{1,M}$ are identically equal on $\mathcal{P}_M^i$, then their difference is identically equal to zero on $\mathcal{P}_M^i$ and the results follows. Otherwise, take $z_0 \in \mathcal{P}_M^i$ such that

$$a := \left(\widetilde{\psi}_{1,M} - \psi_{1,M}\right)\left(z_0\right) \neq 0 \ .$$

By definition of $\mathcal{Z}_M$ and $\mathcal{P}_M$, there exists $s_0 \in M$ such that for any $z_1 \in \mathcal{P}_M^i$,

$$\left(s_0 - s_M\right)\left(z_1\right) = \left(s_0 - s_M\right)\left(z_0\right) \neq 0 \ . \tag{2.58}$$

Hence,

$$\left(\widetilde{\psi}_{1,M} - \psi_{1,M}\right)\left(z_0\right)\left(s_0 - s_M\right)\left(z_0\right) = a\left(s_0 - s_M\right)\left(z_0\right) \neq 0$$

and by Lemma 2.1, it holds

$$\psi_0^{t_{s_0,\lambda}} - \widetilde{\psi}_0^{t_{s_0,\lambda}} \sim \lambda a\left(s_0 - s_M\right)\left(z_0\right) \text{ as } \lambda \to 0 \ . \tag{2.59}$$

Let $z_1 \in \mathcal{P}_M^i$ be fixed. We have

$$\psi_0^{t_{s_0,\lambda}} - \widetilde{\psi}_0^{t_{s_0,\lambda}} = \lambda \left(\widetilde{\psi}_{1,M}(z_1) - \psi_{1,M}(z_1)\right)(s_0 - s_M)(z_1) + \widetilde{\psi}_2\left(\widetilde{\psi}_{3,M}(z_1)(t_{s_0,\lambda} - s_M)(z_1)\right) - \psi_2\left(\psi_{3,M}(z_1)(t_{s_0,\lambda} - s_M)(z_1)\right) .$$

By (2.58), (2.59) and properties (2.51) and (2.52) of Lemma 2.1 applied on z_1, it ensures that

$$\left(\widetilde{\psi}_{1,M}(z_1) - \psi_{1,M}(z_1)\right) = a ,$$

and the first part of Proposition 2.2 is proved.
Take now a function $s_0 \in M$, if there exists $z_0 \in M$ such that $(s_0 - s_M)(z_0) = 0$, then $\psi_0^{t_{s_0,\lambda}}$ is uniquely defined for any $\lambda \in [0,1]$ and so $\psi_0^{t_{s_0,\lambda}} = \widetilde{\psi}_0^{t_{s_0,\lambda}}$. We then have, for any $z \in \mathcal{Z}$ and any $\lambda \in [0,1]$,

$$\lambda \left(\widetilde{\psi}_{1,M}(z) - \psi_{1,M}(z)\right)(s_0 - s_M)(z) = \widetilde{\psi}_2\left(\widetilde{\psi}_{3,M}(z)(t_{s_0,\lambda} - s_M)(z)\right) - \psi_2\left(\psi_{3,M}(z)(t_{s_0,\lambda} - s_M)(z)\right) . \tag{2.60}$$

By dividing each side of equality (2.60) by $\lambda \neq 0$, we get for any $z \in \mathcal{Z}$,

$$\left(\widetilde{\psi}_{1,M}(z) - \psi_{1,M}(z)\right)(s_0 - s_M)(z) = 0 ,$$

since by Lemma 2.1 the right-hand side of (2.60) divided by λ tends to zero as λ tends to zero. We have shown that if $s_0 \in M$ is such that there exists $z_0 \in M$ such that $(s_0 - s_M)(z_0) = 0$, then

$$\widetilde{\psi}_{1,M}(z)(s - s_M)(z) = \psi_{1,M}(z)(s - s_M)(z) \tag{2.61}$$

for any $z \in \mathcal{Z}$. Now, if there exists $z_0 \in M$ such that

$$\left(\widetilde{\psi}_{1,M}(z_0) - \psi_{1,M}(z_0)\right)(s_0 - s_M)(z_0) \neq 0$$

then by the same type of arguments than above in the proof, we have

$$\psi_0^{t_{s_0,\lambda}} - \widetilde{\psi}_0^{t_{s_0,\lambda}} \sim \lambda \left(\widetilde{\psi}_{1,M}(z_0) - \psi_{1,M}(z_0)\right)(s_0 - s_M)(z_0) \text{ as } \lambda \to 0 ,$$

which, by the use of Lemma 2.1, allows to conclude that for any $z \in \mathcal{Z}$,

$$\left(\widetilde{\psi}_{1,M}(z_0) - \psi_{1,M}(z)\right)(s_0 - s_M)(z) = \left(\widetilde{\psi}_{1,M}(z_0) - \psi_{1,M}(z_0)\right)(s_0 - s_M)(z_0) \tag{2.62}$$

and so, by (2.61) and (2.62), we can conclude that the second part of Proposition 2.2 is proved. Now, if $\mathcal{Z} \backslash \mathcal{Z}_M \neq \varnothing$, let us show that $\psi_{1,M}$ and $\widetilde{\psi}_{1,M}$ are identically equal on $\mathcal{Z}_M$. We have that in this case, the constants terms are uniquely defined, given by (2.49), and so for any $s \in M$,

$$\psi_0^s = \widetilde{\psi}_0^s . \tag{2.63}$$

Take $z_0 \in \mathcal{Z}_M$ and assume that $\widetilde{\psi}_{1,M}(z_0) \neq \psi_{1,M}(z_0)$. By definition of $\mathcal{Z}_M$, there exists $s_0 \in M$ such that

$$(s_0 - s_M)(z_0) \neq 0 ,$$

and so

$$\left(\widetilde{\psi}_{1,M}(z_0) - \psi_{1,M}(z_0)\right)(s_0 - s_M)(z_0) \neq 0 . \tag{2.64}$$

By Lemma 2.1 and identity (2.63), we thus have

$$0 = \psi_0^{t_{s_0,\lambda}} - \widetilde{\psi}_0^{t_{s_0,\lambda}} \sim \lambda \left(\widetilde{\psi}_{1,M}(z_0) - \psi_{1,M}(z_0)\right)(s_0 - s_M)(z_0) \text{ as } \lambda \to 0 ,$$

which is in contraction with (2.64). The third part of Proposition 2.2 is thus proved.
Finally, if $\mathcal{Z} = \mathcal{Z}_M$, we assume that there exist $(i,j) \in I^2$ and $(a,b) \in \mathbb{R}_* \times \mathbb{R}_*$ such that

$$a = \widetilde{\psi}_{1,M}(z_i) - \psi_{1,M}(z_i) \neq 0 \text{ and } b = \widetilde{\psi}_{1,M}(z_j) - \psi_{1,M}(z_j) \neq 0$$

for any $(z_i, z_j) \in \mathcal{P}_M^i \times \mathcal{P}_M^j$. If we show that this is not possible then Proposition 2.2 follows. Let $(z_i, z_j) \in \mathcal{P}_M^i \times \mathcal{P}_M^j$ be fixed. By definition of the discriminative partition $\mathcal{P}_M$, there exists $s_0 \in M$, such that

$$a(s_0 - s_M)(z_i) \neq b(s_0 - s_M)(z_j) \ . \tag{2.65}$$

Assume without loss of generality that $a(s_0 - s_M)(z_i) \neq 0$. By Lemma 2.1 we have,

$$\psi_0^{t_{s_0,\lambda}} - \widetilde{\psi}_0^{t_{s_0,\lambda}} \sim \lambda a(s_0 - s_M)(z_i) \text{ as } \lambda \to 0 \ . \tag{2.66}$$

We also have

$$\psi_0^{t_{s_0,\lambda}} - \widetilde{\psi}_0^{t_{s_0,\lambda}} = \lambda b(s_0 - s_M)(z_j) + \widetilde{\psi}_2\left(\widetilde{\psi}_{3,M}(z_j)(t_{s_0,\lambda} - s_M)(z_j)\right) - \psi_2(\psi_{3,M}(z_j)(t_{s_0,\lambda} - s_M)(z_j)) \ .$$

By (2.66) and properties (2.51) and (2.52) of Lemma 2.1 applied on z_i, it ensures that

$$a(s_0 - s_M)(z_i) = b(s_0 - s_M)(z_j) \ ,$$

and the contradiction follows by (2.66), which concludes the proof of Proposition 2.2. ■

Unicity of ψ_2 and $\psi_{3,M}$:

Proposition 2.3 *Assume that the model M is star-shaped at s_M. If there exists $z_0 \in \mathcal{Z}$ and a function $s_0 \in M$ such that*

$$\psi_{3,M}(z_0)(s_0 - s_M)(z_0) \neq 0$$

and for any $\lambda \in (0,1]$, $\psi_0^{t_{s_0,\lambda}}$ is uniquely defined, then the function ψ_2 is defined up to multiplicative constant locally around zero. Moreover, $\psi_{3,M}(z_0)(s_0 - s_M)(z_0)$ is defined up to a multiplicative constant and

$$\psi_2(\lambda \psi_{3,M}(z_0)(s_0 - s_M)(z_0))$$

is uniquely defined locally around zero in λ.

Proof. With the notations of Proposition 2.3, we have for any $\lambda \in (0,1]$,

$$Kt_{s_0,\lambda}(z_0) - Ks_M(z_0) - \psi_0^{t_{s_0,\lambda}} + \psi_{1,M}(z_0)(t_{s_0,\lambda} - s_M)(z_0) = \psi_2(\psi_{3,M}(z_0)(t_{s_0,\lambda} - s_M)(z_0))$$

and so, by definition of $t_{s_0,\lambda}$, we can write

$$\forall \lambda \in (0,1], \ Kt_{s_0,\lambda}(z_0) - Ks_M(z_0) - \psi_0^{t_{s_0,\lambda}} + \lambda \psi_{1,M}(z_0)(s_0 - s_M)(z_0) = \psi_2(\lambda \psi_{3,M}(z_0)(s_0 - s_M)(z_0)) \ . \tag{2.67}$$

Now, we see that the left-hand side of formula (2.67) is uniquely defined for any $\lambda \in (0,1]$. Indeed $Kt_{s_0,\lambda}(z_0) - Ks_M(z_0) - \psi_0^{t_{s_0,\lambda}}$ is uniquely defined and by dividing by λ in each side of (2.67), we see that $\psi_{1,M}(z_0)(s_0 - s_M)(z_0)$ is also uniquely uniquely defined, since by Lemma 2.1, the right-right hand side of (2.67) is of order at most λ^2. Proposition 2.3 then follows. ■

Remark 2.2 *Assume that there exists $z_0 \in \mathcal{Z}$ and $s_0 \in M$ such that*

$$\psi_{3,M}(z_0)(s_0 - s_M)(z_0) \neq 0$$

and

$$\widetilde{\psi}_{3,M}(z_0)(s_0 - s_M)(z_0) \neq 0 \ .$$

If

$$\psi_2(x_1) = \widetilde{\psi}_2(x_1)$$

for some $x_1 \in \left[-A_2 \wedge \tilde{A}_2, A_2 \wedge \tilde{A}_2\right] \setminus \{0\} \subset \mathcal{D}_2 \bigcap \widetilde{\mathcal{D}}_2$, *where* $\mathring{\mathcal{D}}_2$ *and* $\overset{\circ}{\widetilde{\mathcal{D}}}_2$ *are both connected sets, and if* ψ_2 *and* $\widetilde{\psi}_2$ *are analytical on* $\mathring{\mathcal{D}}_2$ *and* $\overset{\circ}{\widetilde{\mathcal{D}}}_2$ *respectively, then, by Proposition 2.3, we see that there exists a function* $\overline{\psi}_2$ *defined on* $\mathcal{D}_2 \bigcup \widetilde{\mathcal{D}}_2$ *such that*

$$\overline{\psi}_2 \equiv \psi_2 \ \textit{on} \ \mathcal{D}_2$$

and

$$\overline{\psi}_2 \equiv \widetilde{\psi}_2 \ \textit{on} \ \widetilde{\mathcal{D}}_2 \ .$$

This proves the uniqueness of ψ_2 *if it is an analytical function on a connected set* $\mathring{\mathcal{D}}_2$ *and if its value on another point than zero - that exists since* $0 \in \mathring{\mathcal{D}}_2$ *- is fixed. Notice that in the three examples of Section 2.2.2,* $\mathring{\mathcal{D}}_2$ *is indeed a connected set and* ψ_2 *is analytical on it.*

Chapitre 3

Optimal upper and lower bounds for the excess risks in heteroscedastic bounded regression

3.1 Introduction

This chapter is devoted to least-squares estimation of a regression function on a finite dimensional linear model. We derive sharp upper and lower bounds in probability for the true and empirical excess risks of the least-squares estimator. We only focus on the "stochastic" parts of the excess risks and we do not discuss on the possible behaviors of the bias of the model, neither on the trade-off that can be achieved between the bias and the variance terms, as we further study model selection procedures related to least-squares regression in Chapter 4. However, our framework is closely related to the method of sieves and particularly to the work of Birgé and Massart [25]. The leading idea of the sieve method is to replace a complicated set of parameters by a more tractable one having good approximation properties, an idea that goes back to Cencov [32], considering orthogonal series for density estimation, and to Le Cam [49] where the author investigate the relationship between the metric structure of the parameter space and the rate of optimal estimators, see also Le Cam [50] Section 16.5 and Le Cam and Yang [51] Section 6.5. Since the formalization of the sieve method by Grenander [40], many authors have considered this method for MLEs or more general M-estimators. Inspired by a work of van de Geer [78] in regression, Birgé and Massart [24] proposed to study minimum of contrast estimation on general parameter spaces under entropy with bracketing conditions, and proved that sub-optimality of M-estimators can happen when the parameter space is too large. The entropy with bracketing covering property has then been a central tool for studying minimum contrast estimation on general sieves in Shen and Wong [66], Wong and Shen [89] and van de Geer [79]. Van de Geer [80] more recently considers M-estimation with convex loss functions, a situation that allows to "localize" the problem to a small neighborhood in the parameter space. In a series of papers that started with Stone [69], Stone extensively studied log-spline density estimation and spline regression, see [68], [70], [71] and Stone and Kooperberg [48].

Birgé and Massart [25] introduced metric properties on the sieves relating the L_2-structure to the L_∞-structure, and which involve covering numbers related to both L_2 and L_∞ norms. These metric conditions are satisfied for linear sieves commonly used in practice, such as Fourier expansions, piecewise polynomials and wavelet expansions, but also for non-linear sieves, which can have better approximation properties, and that include finite linear combinations of D sigmoidal functions related to neural networks, see also Barron [14], and histograms generated by any partition on $[0,1]$ into D subintervals. Birgé and Massart [25] pointed out that the use

of covering numbers, even in the case of linear sieves, is quite natural since linearity is lost on the contrasted functions for a non-linear contrast such as in the regression and maximum likelihood estimation contexts. This allows them to derive sharp exponential bounds and rates of convergence for the excess risk on such sieves, using in particular a Talagrand's concentration inequality for the supremum of the empirical process.

The starting point of our method is to remark that the least-squares contrast in regression is a special case of regular contrast in the sense of Chapter 2 and can thus be expanded to the sum of a linear part and a quadratic part. This allows us to recover some linearity on the contrasted function and avoid the use of entropy methods to control the empirical process on a linear model. The gain is that we achieve optimal rates of convergence for the true and empirical excess risks with exact constants, for models of reasonable dimension. In our study, the metric properties defined by Birgé and Massart in [25] play a center role, in particular the notion of localized basis. In addition, we point out the importance of the behavior in sup-norm of the least-squares estimator and we have to assume its consistency in sup-norm towards the linear projection of the regression function onto the model. We show that such a condition is satisfied by histograms and piecewise polynomial models when they are endowed with a localized basis structure, which corresponds in that case to a lower regularity assumption on the considered partition. By doing so, we recover some recent results of Arlot and Massart [10] on the empirical and true excess risk for least-squares estimator on histogram models, and extend them to the case of piecewise polynomials.

Altough we do not make an explicit use of the margin conditions that can hold in the context regular contrast and more especially in the context of bounded regression, see Chapter 2, this property also connects our work with the statistical learning theory. The margin conditions were first introduced by Mammen and Tsybakov [60] in the context of discrimination analysis. They allow to get faster rates of convergence than the pioneering bounds of Vapnik and Červonenkis, see [83] and [82], using "localization" techniques. Under entropy with bracketing conditions, Tsybakov [75] shows some fast rates in the binary classification setting, and these results have been recovered and extended by Massart and Nédélec [62], Koltchinskii [44] and by Giné and Koltchinskii [39], where the authors also give asymptotic results for ratio type empirical processes. The obtained bounds are proved to be optimal in a minimax sense in [62], up to a logarithmic factor shown by Massart and Nédélec to be unavoidable for "rich" VC-classes. This analysis is refined in [39] by the use of localized $L_2(P)$-envelopes of the models, allowing to remove the logarithmic factor in good cases.

The main tools in [62], [44] and [39] are Talagrand's type concentration inequalities for the supremum of the empirical process and the *slicing* or *pealing* technique through the use of ratio type empirical processes. The slicing technique consists in considering subsets of the model, called the slices, and that are localized in terms of excess risk, a quantity that is related to the variance of the empirical process through margin conditions. Our method of proof may be viewed as a variant of the technique of slicing that allows to avoid the use of ratio type empirical processes, where in general sharp constants are lost due to the use of chaining techniques. The very first lines of our proofs differ from those of [62], [44] and [39], and permit in particular to relate both upper and lower bounds for the excess risks of the M-estimator to the behavior of the empirical process indexed by contrasted functions on localized slices of excess risk. This rewriting of the problem of upper and lower bounds for the excess risks is closely related to the work of Bartlett and Mendelson [19], were a "direct" approach of the empirical minimization algorithm is proposed, and proved to lead to more accurate bounds than the traditional "structural" approach developed in [62], [39] or [44].

The chapter is organized as follows. We present the statistical framework in Section 3.2 where we show in particular the regularity of the least-squares regression contrast. We then derive general results for models of reasonable dimensions and also for small models in Section 3.3. General results are then applied in the case of histograms and piecewise polynomials in

Section 3.4 and 3.5 respectively, where explicit rates of convergence in sup-norm are derived. The proofs are postponed to the end of the chapter.

3.2 Regression framework and notations

3.2.1 Least-squares estimator

Let $(\mathcal{X}, \mathcal{T}_{\mathcal{X}})$ be a measurable space and set $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$. We assume that $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, $i \in \{1, ..., n\}$ are n i.i.d. observations with law P. The marginal law of X_i is denoted by P^X. We assume that the data satisfy the following relation

$$Y_i = s_* (X_i) + \sigma (X_i) \varepsilon_i \ , \tag{3.1}$$

where $s_* \in L_2 \left(P^X\right)$, ε_i are i.i.d. random variables with mean 0 and variance 1 conditionally to X_i and $\sigma : \mathcal{X} \longrightarrow \mathbb{R}$ is an heteroscedastic noise level. A generic random variable of law P, independent of $(\xi_1, ..., \xi_n)$, is denoted by $\xi = (X, Y)$.
Hence, s_* is the regression function of Y with respect to X, that we want to estimate. Given a finite dimensional linear vector space M, we denote by s_M the linear projection of s_* onto M in $L_2 \left(P^X\right)$ and by D the linear dimension of the model M.
We consider on M a least-squares estimator s_n (possibly non unique), defined as follows

$$s_n \in \arg\min_{s \in M} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - s (X_i))^2 \right\} \ . \tag{3.2}$$

So, if we denote by

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$$

the empirical distribution of the data and by $K : L_2 \left(P^X\right) \longrightarrow L_1 (P)$ the least-squares contrast, defined by

$$K (s) = (x, y) \in \mathcal{Z} \to (y - s (x))^2 , \quad s \in L_2 \left(P^X\right)$$

we then remark that s_n belongs to the general class of M-estimators, as it satisfies

$$s_n \in \arg\min_{s \in M} \{P_n (K (s))\} \ . \tag{3.3}$$

3.2.2 Excess risk and contrast

As defined in (3.3), s_n is the empirical risk minimizer of the least-squares contrast. The regression function s_* can be defined as the minimizer in $L_2 \left(P^X\right)$ of the mean of the contrast over the unknown law P,

$$s_* = \arg\min_{s \in L_2(P^X)} PK (s) \ ,$$

where

$$PK (s) = P (Ks) = PKs = \mathbb{E} [K (s) (X, Y)] = \mathbb{E} \left[(Y - s (X))^2 \right]$$

is called the risk of the function s. In particular we have $PKs_* = \mathbb{E} \left[\sigma^2 (X)\right]$. We first notice that for any $s \in L_2 \left(P^X\right)$, if we denote by

$$\|s\|_2 = \left(\int_{\mathcal{X}} s^2 dP^X \right)^{1/2}$$

its quadratic norm, then we have, by (3.1) above,

$$\begin{aligned} PKs - PKs_* &= P(Ks - Ks_*) \\ &= \mathbb{E}\left[(Y - s(X))^2 - (Y - s_*(X))^2\right] \\ &= \mathbb{E}\left[(s_* - s)(X)(2(Y - s_*(X)) + (s_* - s)(X))\right] \\ &= \mathbb{E}\left[(s_* - s)^2(X)\right] + 2\mathbb{E}\left[(s_* - s)(X)\underbrace{\mathbb{E}\left[Y - s_*(X)\,|X\right]}_{=0}\right] \\ &= \|s - s_*\|_2^2 \geq 0\ , \end{aligned}$$

and $PKs - PKs_*$ is called the excess risk of s. So if we denote by s_M the linear projection of s_* onto M in $L_2\left(P^X\right)$, we have

$$PKs_M - PKs_* = \inf_{s \in M}\left\{PKs - PKs_*\right\}\ , \tag{3.4}$$

and for all $s \in M$

$$P^X\left(s \cdot (s_M - s_*)\right) = 0\ . \tag{3.5}$$

From (3.4), we deduce that

$$s_M = \arg\min_{s \in M} PK(s)\ .$$

Our goal is to study the performance of the least-squares estimator, that we measure by its excess risk. So we are mainly interested by the random quantity $P(Ks_n(M) - Ks_*)$. Moreover, as we can write

$$P(Ks_n(M) - Ks_*) = P(Ks_n(M) - Ks_M) + P(Ks_M - Ks_*)$$

we naturally focus on the quantity

$$P(Ks_n(M) - Ks_M) \geq 0$$

that we want to upper and lower bound in probability. Abusively we will often call this last quantity the excess risk of the estimator on M or the true excess risk of $s_n(M)$, in opposition to the empirical excess risk for which the expectation is taken over the empirical measure,

$$P_n(Ks_M - Ks_n(M)) \geq 0.$$

The following lemma establishes the key expansion of the regression contrast around s_M on M. This expansion exhibits a linear part and a quadratic part. This is an example of what we call more generally a regular contrast, see Chapter 2.

Lemma 3.1 *We have, for every $z = (x, y) \in \mathcal{Z}$,*

$$(Ks)(z) - (Ks_M)(z) = \psi_{1,M}(z)(s - s_M)(x) + \psi_2((s - s_M)(x)) \tag{3.6}$$

with $\psi_{1,M}(z) = -2(y - s_M(x))$ and $\psi_2(t) = t^2$, for all $t \in \mathbb{R}$. Moreover, for all $s \in M$,

$$P(\psi_{1,M} \cdot s) = 0\ . \tag{3.7}$$

Proof. Start with

$$\begin{aligned} &(Ks)(z) - (Ks_M)(z) \\ &= (y - s(x))^2 - (y - s_M(x))^2 \\ &= ((s - s_M)(x))((s - s_M)(x) - 2(y - s_M(x))) \\ &= -2(y - s_M(x))((s - s_M)(x)) + ((s - s_M)(x))^2\ , \end{aligned}$$

which gives (3.6). Moreover, observe that for any $s \in M$,

$$P\left(\psi_{1,M} \cdot s\right) = -2\mathbb{E}\left[\left(Y - s_*\left(X\right)\right) s\left(X\right)\right] + 2\mathbb{E}\left[s\left(X\right)\left(s_M - s_*\right)\left(X\right)\right] \ . \tag{3.8}$$

We have

$$\mathbb{E}\left[\left(Y - s_*\left(X\right)\right) s\left(X\right)\right] = \mathbb{E}\left[\underbrace{\mathbb{E}\left[\left(Y - s_*\left(X\right)\right)|X\right]}_{=0} s\left(X\right)\right] = 0 \ . \tag{3.9}$$

and, by (3.5),

$$\mathbb{E}\left[s\left(X\right)\left(s_M - s_*\right)\left(X\right)\right] = P^X\left(s \cdot \left(s_M - s_*\right)\right) = 0 \ . \tag{3.10}$$

Combining (3.8), (3.9) and (3.10) we get that for any $s \in M$, $P\left(\psi_{1,M} \cdot s\right) = 0$. This concludes the proof. ■

3.3 True and empirical excess risk bounds

In this section, we show that under assumptions that extend a previous work of Arlot and Massart [10], the true excess risk is equivalent to the empirical one for models of reasonable dimension, which is a keystone to prove the slope phenomenon that we expose in Chapter 4. More precisely, we assume that M is a linear model with a localized basis in $L_2\left(P\right)$ and that the least-squares estimator is consistent in sup-norm towards the linear projection s_M on M of the target s_* when the dimension of the model is not too heavy. This is a natural generalization of the case of histograms studied by Arlot and Massart in [10], since the assumption of lower regularity of the partitions made in their work indeed provides the histograms with a structure of localized basis in $L_2\left(P\right)$, see Lemma 3.2. We further show in Lemma 3.3 that the assumption of consistency is satisfied for histograms.

3.3.1 Main assumptions

We turn now to the statement of some assumptions that will be needed to derive our results in Section 3.3.2. These assumptions will be further discussed in Section 3.3.3.

Boundedness assumptions :

- (**H1**) The data and the linear projection of the target onto M are bounded : a positive finite constant A exists such that

$$|Y_i| \leq A \ a.s. \tag{3.11}$$

and

$$\left\|s_M\right\|_\infty \leq A \ . \tag{3.12}$$

Hence, from (**H1**) we deduce that

$$\left\|s_*\right\|_\infty = \left\|\mathbb{E}\left[Y|X = \cdot\right]\right\|_\infty \leq A \tag{3.13}$$

and that there exists a constant $\sigma_{\max} > 0$ such that

$$\sigma^2\left(X_i\right) \leq \sigma_{\max}^2 \leq A^2 \ \ a.s. \tag{3.14}$$

Moreover, as $\psi_{1,M}\left(z\right) = -2\left(y - s_M\left(x\right)\right)$ for all $z = (x, y) \in \mathcal{Z}$, we also deduce that

$$\left|\psi_{1,M}\left(X_i, Y_i\right)\right| \leq 4A \ \ a.s. \tag{3.15}$$

- (**H2**) The heteroscedastic noise level σ is uniformly bounded from below : a positive finite constant $\sigma_{\min}$ exists such that

$$0 < \sigma_{\min} \leq \sigma(X_i) \quad a.s.$$

Models with localized basis in $L_2(P^X)$:

Let us define a function Ψ_M on $\mathcal{X}$, that we call the unit envelope of M, such that

$$\Psi_M(x) = \frac{1}{\sqrt{D}} \sup_{s \in M, \|s\|_2 \leq 1} |s(x)| \ . \tag{3.16}$$

As M is a finite dimensional real vector space, the supremum in (3.16) can also be taken over a countable subset of M, so Ψ_M is a measurable function.

- (**H3**) The unit envelope of M is uniformly bounded on $\mathcal{X}$: a positive constant $A_{3,M}$ exists such that

$$\|\Psi_M\|_\infty \leq A_{3,M} < \infty \ .$$

The following assumption is stronger than (**H3**).

- (**H4**) Existence of a localized basis in $(M, \|\cdot\|_2)$: there exists an orthonormal basis $\varphi = (\varphi_k)_{k=1}^D$ in $(M, \|\cdot\|_2)$ that satisfies, for a positive constant $r_M(\varphi)$ and all $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$,

$$\left\| \sum_{k=1}^D \beta_k \varphi_k \right\|_\infty \leq r_M(\varphi) \sqrt{D} \, |\beta|_\infty \ ,$$

where $|\beta|_\infty = \max\{|\beta_k| \, ; k \in \{1, ..., D\}\}$ is the sup-norm of the D-dimensional vector β.

Remark 3.1 *(**H4**) implies (**H3**) and in that case $A_{3,M} = r_M(\varphi)$ is convenient.*

The assumption of consistency in sup-norm :

In order to handle second order terms in the expansion of the contrast (3.6) we assume that the least-squares estimator is consistent for the sup-norm on the space $\mathcal{X}$. More precisely, this requirement can be stated as follows.

- (**H5**) Assumption of consistency in sup-norm **:** for any $A_+ > 0$, if M is a model of dimension D satisfying

$$D \leq A_+ \frac{n}{(\ln n)^2} \ ,$$

then for every $\alpha > 0$, we can find a positive integer n_1 and a positive constant A_{cons} satisfying the following property : there exists $R_{n,D,\alpha} > 0$ depending on D, n and α, such that

$$R_{n,D,\alpha} \leq \frac{A_{cons}}{\sqrt{\ln n}} \tag{3.17}$$

and by setting

$$\Omega_{\infty,\alpha} = \{\|s_n - s_M\|_\infty \leq R_{n,D,\alpha}\} \ , \tag{3.18}$$

it holds for all $n \geq n_1$,

$$\mathbb{P}[\Omega_{\infty,\alpha}] \geq 1 - n^{-\alpha} \ . \tag{3.19}$$

3.3.2 Theorems

We state here the general results of this chapter, that will be applied in Section 3.4 and 3.5 in the case of piecewise constant functions and piecewise polynomials respectively.

Theorem 3.1 *Let $A_+, A_-, \alpha > 0$ and let M be a linear model of finite dimension D. Assume that (**H1**), (**H2**), (**H4**) and (**H5**) hold and take $\varphi = (\varphi_k)_{k=1}^D$ an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying (**H4**). If it holds*

$$A_- (\ln n)^2 \leq D \leq A_+ \frac{n}{(\ln n)^2} \ , \tag{3.20}$$

*then a positive finite constant A_0 exists, only depending on α, A_- and on the constants $A, \sigma_{\min}, r_M(\varphi)$ defined in the assumptions (**H1**), (**H2**) and (**H4**) respectively, such that by setting*

$$\varepsilon_n = A_0 \max\left\{ \left(\frac{\ln n}{D}\right)^{1/4}, \ \left(\frac{D \ln n}{n}\right)^{1/4}, \ \sqrt{R_{n,D,\alpha}} \right\} \ , \tag{3.21}$$

we have for all $n \geq n_0(A_-, A_+, A, A_{cons}, r_M(\varphi), \sigma_{\min}, n_1, \alpha)$,

$$\mathbb{P}\left[P(Ks_n - Ks_M) \geq (1 - \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-\alpha} \ , \tag{3.22}$$

$$\mathbb{P}\left[P(Ks_n - Ks_M) \leq (1 + \varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-\alpha} \ , \tag{3.23}$$

$$\mathbb{P}\left[P_n(Ks_M - Ks_n) \geq \left(1 - \varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 2n^{-\alpha} \ , \tag{3.24}$$

$$\mathbb{P}\left[P_n(Ks_M - Ks_n) \leq \left(1 + \varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 3n^{-\alpha} \ , \tag{3.25}$$

*where $\mathcal{K}_{1,M}^2 = \frac{1}{D} \sum_{k=1}^D \mathrm{Var}(\psi_{1,M} \cdot \varphi_k)$. In addition, when (**H5**) does not hold, but (**H1**), (**H2**) and (**H4**) hold, we still have for all $n \geq n_0(A_-, A_+, A, r_M(\varphi), \sigma_{\min}, \alpha)$,*

$$\mathbb{P}\left(P_n(Ks_M - Ks_n) \geq \left(1 - A_0 \max\left\{ \sqrt{\frac{\ln n}{D}}, \sqrt{\frac{D \ln n}{n}} \right\}\right) \frac{D}{4n} \mathcal{K}_{1,M}^2 \right) \geq 1 - 2n^{-\alpha} \ . \tag{3.26}$$

In Theorem 3.1 above, we achieve sharp upper and lower bounds for the true and empirical excess risks on M. They are optimal at the first order since the leading constants are equal for upper and lower bounds. Moreover, Theorem 3.1 states the equivalence with high probability of the true and empirical excess risks for models of reasonable dimensions, which is the starting point of the slope heuristics as explained in Chapter 4. We can notice that second orders are smaller for the empirical excess risk than for the true one. Indeed, when normalized by the first order, the deviations of the empirical excess risk are square of the deviations of the true one. Our bounds also give another evidence of the concentration phenomenon of the empirical excess risk exhibited by Boucheron and Massart [27] in the slightly different context of M-estimation with bounded contrast where some margin condition hold. Notice that considering the lower bound of the empirical excess risk given in (3.26), we do not need to assume the consistency of the least-squares estimator s_n towards the linear projection s_M.

We turn now to upper bounds in probability for the true and empirical excess risks on models with possibly small dimensions. We do not achieve sharp or explicit constants in the rates of convergence and in fact, information given by Theorem 3.2 below suffices to our needs, as we use it in the proofs of the theorems stated in Chapter 4, when we have to control model selection procedures for small models.

Theorem 3.2 *Let $\alpha, A_+ > 0$ be fixed and let M be a linear model of finite dimension*

$$1 \leq D \leq A_+ \frac{n}{(\ln n)^2} \ .$$

*Assume that assumptions (**H1**), (**H3**) and (**H5**) hold. Then a positive constant A_u exists, only depending on $A, A_{cons}, A_{3,M}$ and α, such that for all $n \geq n_0 (A_{cons}, n_1)$,*

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \tag{3.27}$$

and

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \ . \tag{3.28}$$

Notice that on contrary to the situation of Theorem 3.1, we do not assume that (**H2**) hold. This assumption states that the noise level is uniformly bounded away from zero over the space $\mathcal{X}$, and allows in Theorem 3.1 to derive lower bounds for the true and empirical excess risks, as well as to achieve sharp constants in the deviation bounds for models of reasonable dimensions. In Theorem 3.2, we just derive upper bounds and assumption (**H2**) is not needed. The price to pay is that constants in the rates of convergence derived in (3.27) and (3.28) are possibly larger than the corresponding ones of Theorem 3.1, but our results still hold true for small models. Moreover, in the case of models with reasonable dimensions, that is dimensions satisfying assumption (3.20) of Theorem 3.1, the rate of decay is preserved compared to Theorem 3.1 and is proportional to D/n.
The proofs of the above theorems can be found in Section 3.6.3.

3.3.3 Some additional comments

Let us first comment on the assumptions given in Section 3.3.1. Assumptions (3.11) and (**H2**) are rather mild and can also be found in the work of Arlot and Massart [10] related to the case of histograms, where they are respectively denoted by (**Ab**) and (**An**). The histogram case will be further commented in Section 3.4.3.
In assumption (**H4**) we require that the model M is provided with an orthonormal localized basis in $L_2\left(P^X\right)$. This property is convenient when dealing with the L_∞-structure on the model, and this allows us to control the sup-norm of the functions in the model by the sup-norm of the vector of their coordinates in the localized basis. For examples of models with localized basis, and their use in a model selection framework, we refer for instance to Section 7.4.2 of Massart [61], where it is shown that models of histograms, piecewise polynomials and compactly supported wavelets are typical examples of models with localized basis for the $L_2\left(\text{Leb}\right)$ structure, considering that $\mathcal{X} \subset \mathbb{R}^k$. In Sections 3.4 and 3.5, we show that models of piecewise constant and piecewise polynomials respectively can also have a localized basis for the $L_2\left(P^X\right)$ structure, under rather mild assumptions on P^X. Assumption (**H4**) is needed in Theorem 3.1, whereas in Theorem 3.2 we only use the weaker assumption (**H3**) on the unit envelope of the model M, relating the L_2-structure of the model to the L_∞-structure. In fact, assumption (**H4**) allows us in the proof of Theorem 3.1 to achieve sharp lower bounds for the quantities of interest, whereas in Theorem 3.2 we only give upper bounds in the case of small models.
We ask in assumption (**H5**) that the M-estimator is consistent towards the linear projection s_M of s_* onto the model M, at a rate at least better than $(\ln n)^{-1/2}$. This can be considered as a rather strong assumption, but it is essential for our methodology. Moreover, we show in Sections 3.4 and 3.5 that this assumption is satisfied under mild conditions for histogram models and models of piecewise polynomials respectively, both at the rate

$$R_{n,D,\alpha} \propto \sqrt{\frac{D \ln n}{n}} \ .$$

Secondly, let us comment on the rates of convergence given in Theorem 3.1 for models of reasonable dimensions. As we can see in Theorem 3.1, the rate of estimation in a fixed model M of reasonable dimension is determined at the first order by a key quantity that relates the structure of the model to the unknown law P of data. We call this quantity the **complexity** of the model M and we denote it by $\mathcal{C}_M$. More precisely, let us define

$$\mathcal{C}_M = \frac{1}{4} D \times \mathcal{K}_{1,M}^2$$

where

$$\mathcal{K}_{1,M} = \sqrt{\frac{1}{D} \sum_{k=1}^{D} \mathrm{Var}\left(\psi_{1,M} \cdot \varphi_k\right)}$$

for a localized orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$. Notice that $\mathcal{K}_{1,M}$ is well defined as it does not depend on the choice of the basis $(\varphi_k)_{k=1}^D$. Indeed, since we have $P\left(\psi_{1,M} \cdot \varphi_k\right) = 0$, we deduce that

$$\mathcal{K}_{1,M}^2 = P\left(\psi_{1,M}^2 \cdot \left(\frac{1}{D} \sum_{k=1}^{D} \varphi_k^2\right)\right) .$$

Now observe that, by using Cauchy-Schwarz inequality in Definition (3.16), as pointed out by Birgé and Massart [25], we get

$$\Psi_M^2 = \frac{1}{D} \sum_{k=1}^{D} \varphi_k^2 \tag{3.29}$$

and so

$$\begin{aligned} \mathcal{K}_{1,M}^2 &= P\left(\psi_{1,M}^2 \Psi_M^2\right) \\ &= 4\mathbb{E}\left[\mathbb{E}\left[\left(Y - s_M(X)\right)^2 | X\right] \Psi_M^2(X)\right] \\ &= 4\left(\mathbb{E}\left[\sigma^2(X) \Psi_M^2(X)\right] + \mathbb{E}\left[\left(s_M - s_*\right)^2(X) \Psi_M^2(X)\right]\right) . \end{aligned} \tag{3.30}$$

On the one hand, if we assume (**H1**) then we obtain by elementary computations

$$\mathcal{K}_{1,M} \leq 2\sigma_{\max} + 4A \leq 6A . \tag{3.31}$$

On the other hand, (**H2**) implies

$$\mathcal{K}_{1,M} \geq 2\sigma_{\min} > 0 . \tag{3.32}$$

To fix ideas, let us explicitly compute $\mathcal{K}_{1,M}^2$ in a simple case. Consider homoscedastic regression on a histogram model M, in which the homoscedastic noise level σ is such that

$$\sigma^2(X) = \sigma^2 \quad a.s. ,$$

so that we have

$$\mathbb{E}\left[\sigma^2(X) \Psi_M^2(X)\right] = \sigma^2 \mathbb{E}\left[\Psi_M^2(X)\right] = \sigma^2 .$$

Now, under notations of Lemma 3.2 below,

$$s_M = \sum_{I \in \mathcal{P}} \mathbb{E}\left[Y \varphi_I(X)\right] \varphi_I = \sum_{I \in \mathcal{P}} \mathbb{E}\left[Y | X \in I\right] \mathbf{1}_I ,$$

thus we deduce, by (3.29) and the previous equality, that

$$\begin{aligned}\mathbb{E}\left[(s_M - s_*)^2(X)\,\Psi_M^2(X)\right] &= \frac{1}{|\mathcal{P}|}\sum_{I\in\mathcal{P}}\mathbb{E}\left[(s_M - s_*)^2(X)\,\varphi_I^2(X)\right] \\ &= \frac{1}{|\mathcal{P}|}\sum_{I\in\mathcal{P}}\mathbb{E}\left[(\mathbb{E}[Y|X\in I] - \mathbb{E}[Y|X])^2\,\frac{\mathbf{1}_{X\in I}}{P^X(I)}\right] \\ &= \frac{1}{|\mathcal{P}|}\sum_{I\in\mathcal{P}}\mathbb{E}\left[(\mathbb{E}[Y|X\in I] - \mathbb{E}[Y|X])^2\,|X\in I\right] \\ &= \frac{1}{|\mathcal{P}|}\sum_{I\in\mathcal{P}}\mathbb{V}\left[\mathbb{E}[Y|X]\,|X\in I\right]\ ,\end{aligned}$$

where the conditional variance $\mathbb{V}[U|\mathcal{A}]$ of a variable U with respect to the event $\mathcal{A}$ is defined to be

$$\mathbb{V}[U|\mathcal{A}] := \mathbb{E}\left[(U - \mathbb{E}[U|\mathcal{A}])^2|\mathcal{A}\right] = \mathbb{E}\left[U^2|\mathcal{A}\right] - (\mathbb{E}[U|\mathcal{A}])^2\ .$$

By (3.30), we explicitly get

$$\mathcal{K}_{1,M}^2 = 4\left(\sigma^2 + \frac{1}{|\mathcal{P}|}\sum_{I\in\mathcal{P}}\mathbb{V}\left[\mathbb{E}[Y|X]\,|X\in I\right]\right)\ . \tag{3.33}$$

A careful look at the proof of Theorem 3.1 given in Section 3.6.3 show that condition (**H2**) is only used through the lower bound (3.32), and thus (**H2**) can be replaced by the following slightly more general assumption :

(H2bis) Lower bound on the normalized complexity $\mathcal{K}_{1,M}$: a positive constant $A_{\min}$ exists such that

$$\mathcal{K}_{1,M} \geq A_{\min} > 0\ .$$

When (**H2**) holds, we see from Inequality 3.32 that (**H2bis**) is satisfied with $A_{\min} = 2\sigma_{\min}$. For suitable models we can have for a positive constant A_Ψ^- and for all $x \in \mathcal{X}$,

$$\Psi_M(x) \geq A_\Psi^- > 0\ , \tag{3.34}$$

and this allows to consider vanishing noise level, as we then have by (3.30),

$$\mathcal{K}_{1,M} \geq 2A_\Psi^-\sqrt{\mathbb{E}\left[\sigma^2(X)\right]} = 2A_\Psi^-\left\|\sigma\right\|_2 > 0\ .$$

As we will see in Sections 3.4 and 3.5, Inequality (3.34) can be satisfied for histogram and piecewise polynomial models on a partition achieving some upper regularity assumption with respect to the law P^X .

3.4 The histogram case

In this section, we particularize the results stated in Section 3.3 to the case of piecewise constant functions. We show that under a lower regularity assumption on the considered partition, the assumption (**H4**) of existence of a localized basis in $L_2\left(P^X\right)$ and (**H5**) of consistency in sup-norm of the M-estimator towards the linear projection s_M are satisfied.

3.4.1 Existence of a localized basis

The following lemma states the existence of an orthonormal localized basis for piecewise constant functions in $L_2\left(P^X\right)$, on a partition which is lower-regular for the law P^X.

Lemma 3.2 *Let consider a linear model M of histograms defined on a finite partition $\mathcal{P}$ on $\mathcal{X}$, and write $|\mathcal{P}| = D$ the dimension of M. Moreover, assume that for a positive finite constant $c_{M,P}$,*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0 \ . \tag{3.35}$$

Set, for $I \in \mathcal{P}$,

$$\varphi_I = \left(P^X(I)\right)^{-1/2} \mathbf{1}_I \ .$$

Then the family $(\varphi_I)_{I \in \Lambda_M}$ is an orthonormal basis in $L_2\left(P^X\right)$ and we have,

$$\text{for all } \beta = (\beta_I)_{I \in \mathcal{P}} \in \mathbb{R}^D, \quad \left\| \sum_{I \in \mathcal{P}} \beta_I \varphi_I \right\|_\infty \leq c_{M,P}^{-1} \sqrt{D}\, |\beta|_\infty \ . \tag{3.36}$$

Condition (3.35) can also be found in Arlot and Massart [10] and is named lower regularity of the partition $\mathcal{P}$ for the law P^X. It is easy to see that the lower regularity of the partition is equivalent to the property of localized basis in the case of histograms, i.e. (3.35) is equivalent to (3.36). The proof of Lemma 3.2 is straightforward and can be found in Section 3.6.1.

3.4.2 Rates of convergence in sup-norm

The following lemma allows to derive property (**H5**) for histogram models.

Lemma 3.3 *Consider a linear model M of histograms defined on a finite partition $\mathcal{P}$ of $\mathcal{X}$, and denote by $|\mathcal{P}| = D$ the dimension of M. Assume that Inequality (3.11) holds, that is, a positive constant A exists such that $|Y| \leq A$ a.s. Moreover, assume that for some positive finite constant $c_{M,P}$,*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0 \tag{3.37}$$

and that $D \leq A_+ n (\ln n)^{-2} \leq n$ for some positive finite constant A_+. Then, for any $\alpha > 0$ and for all $n \geq n_0(\alpha, c_{M,P}, A_+)$, there exists an event of probability at least $1 - n^{-\alpha}$ on which s_n exists, is unique and it holds,

$$\|s_n - s_M\|_\infty \leq L_{A_+, A, c_{M,P}, \alpha} \sqrt{\frac{D \ln n}{n}} \ . \tag{3.38}$$

In Lemma 3.3 we thus achieve the convergence in sup-norm of the regressogram s_n towards the linear projection s_M at the rate $\sqrt{D \ln(n)/n}$. It is worth noticing that for a model of histograms satisfying the assumptions of Lemma 3.3, if we set

$$A_{cons} = L_{A, c_{M,P}, \alpha} \sqrt{A_+} \ , \ n_1 = n_0(\alpha, c_{M,P}, A_+) \text{ and } R_{n,D,\alpha} = L_{A_+, A, c_{M,P}, \alpha} \sqrt{\frac{D \ln n}{n}},$$

then Assumption **(H5)** is satisfied. To derive Inequality (3.38), we need to assume that the response variable Y is almost surely bounded and that the considered partition is lower-regular for the law P^X. Hence, we fit again with the framework of [10] and we can thus view the general set of assumptions exposed in Section 3.3.1 as a natural generalization for linear models of the framework developed in [10] in the case of histograms. The proof of Lemma 3.3 can be found in Section 3.6.1.

3.4.3 Bounds for the excess risks

The next theorem is a straightforward application of Lemmas 3.2, 3.3 and Theorems 3.1, 3.2. Indeed, we recover results of Theorems 3.1 and 3.2 for models of histograms, under the lower regularity assumption on the considered partition of the space $\mathcal{X}$ with respect to the unknown law P^X. As seen in Section 3.4.2, we have in that case

$$R_{n,D,\alpha} \propto \sqrt{\frac{D \ln n}{n}} \ .$$

Theorem 3.3 *Given $A_+, A_-, \alpha > 0$, consider a linear model M of histograms defined on a finite partition $\mathcal{P}$ of $\mathcal{X}$, and write $|\mathcal{P}| = D$ the dimension of M. Assume that for some positive finite constant $c_{M,P}$, it holds*

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} P^X(I)} \geq c_{M,P} > 0 \ . \tag{3.39}$$

*If (**H1**) and (**H2**) of Section 3.3.1 are satisfied and if*

$$A_- (\ln n)^2 \leq D \leq A_+ \frac{n}{(\ln n)^2} \ ,$$

then there exists a positive finite constant A_0, only depending on $\alpha, A, \sigma_{\min}, A_-, A_+, c_{M,P}$ such that, by setting

$$\varepsilon_n = A_0 \max\left\{ \left(\frac{\ln n}{D}\right)^{1/4}, \left(\frac{D \ln n}{n}\right)^{1/4} \right\}$$

we have, for all $n \geq n_0(A_-, A_+, A, , \sigma_{\min}, c_{M,P}, \alpha)$,

$$\mathbb{P}\left[(1+\varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq P(Ks_n - Ks_M) \geq (1-\varepsilon_n) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 10 n^{-\alpha} \tag{3.40}$$

and

$$\mathbb{P}\left[\left(1+\varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq P_n(Ks_M - Ks_n) \geq \left(1-\varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5 n^{-\alpha} \ . \tag{3.41}$$

*If (3.39) holds together with (**H1**) and if we assume that*

$$1 \leq D \leq A_+ \frac{n}{(\ln n)^2} \ ,$$

then a positive constant A_u exists, only depending on $A, c_{M,P}, A_+$ and α, such that for all $n \geq n_0(A, c_{M,P}, A_+, \alpha)$,

$$\mathbb{P}\left[P(Ks_n - Ks_M) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3 n^{-\alpha}$$

and

$$\mathbb{P}\left[P_n(Ks_M - Ks_n) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3 n^{-\alpha} \ .$$

As announced before, we recover in Theorem 3.3 the general results of Section 3.3.2 for the case of histograms on a lower-regular partition. Moreover, in the case of histograms, assumption (3.12) which is part of (**H1**) is a straightforward consequence of (3.11). Indeed, we easily see that the projection s_M of the regression function s_* onto the model of piecewise constant functions with respect to $\mathcal{P}$ can be written

$$s_M = \sum_{I \in \mathcal{P}} \mathbb{E}[Y | X \in I] \mathbf{1}_I \ . \tag{3.42}$$

Under (3.11), we have $|\mathbb{E}[Y | X \in I]| \leq \|Y\|_\infty \leq A$ for every $I \in \mathcal{P}$ and we deduce by (3.42) that $\|s_M\|_\infty \leq A$.

3.4.4 Comments

Our bounds in Theorem 3.3 are obtained by following a general methodology as exposed in Chapters 2 and 7. It is then instructive to compare them to the best available results in this special case. Let us compare them to the bounds obtained by Arlot and Massart in [10], in the case of a fixed model. Such results can be found in Proposition 10, 11 and 12 of [10].
The strategy adopted by the authors in this case is as follows. By remarking that easy bounds are available for the mean of the empirical excess risk on histograms since it holds

$$\mathbb{E}\left[P_n\left(Ks_M - Ks_n\right)\right] = \frac{D}{4n}\mathcal{K}_{1,M}^2 \ ,$$

they derive concentration inequalities for the true excess risk and its empirical counterpart to their mean. They further give upper and lower bounds in terms of $\mathbb{E}\left[P_n\left(Ks_M - Ks_n\right)\right]$ for the mean of the true excess risk. The deviations in all these inequalities are made of sums of quantities that can not be compared to ours in a concise manner, as some of them loose compared to our results and some of them gain.
Nevertheless, using our notations, Inequality (34) of Proposition 10 in [10] states that for every $x \geq 0$ there exists an event of probability at least $1 - e^{1-x}$ on which,

$$\begin{aligned} &\left|P_n\left(Ks_M - Ks_n\right) - \mathbb{E}\left[P_n\left(Ks_M - Ks_n\right)\right]\right| \\ &\leq \frac{L}{\sqrt{D_M}}\left[P(Ks_M - Ks_*) + \frac{A^2\mathbb{E}\left[P_n\left(Ks_M - Ks_n\right)\right]}{\sigma_{\min}^2}\left(\sqrt{x} + x\right)\right] \ , \end{aligned} \tag{3.43}$$

for some absolute constant L. We can notice that Inequality (3.43), which is a special case of general concentration inequalities given by Boucheron and Massart [27], involves the bias of the model $P(Ks_M - Ks_*)$. By pointing out that the bias term arises from the use of some margin conditions that are satisfied for bounded regression, we believe that it can be removed from Proposition 10 of [10], since in the case of histograms models for bounded regression, some margin-like conditions hold, that are directly pointed at the linear projection s_M, see Section 2.2.3 of Chapter 2 for a proof of this fact. Apart for the bias term, the deviations of the empirical excess risk are then of the order

$$\frac{\ln\left(n\right)\sqrt{D_M}}{n} \ ,$$

considering the same probability of event than ours, so it becomes significantly better than Inequality (3.41) for large models.
Concentration inequalities for the true excess risk given in Proposition 11 of [10] give a magnitude of deviations that is again smaller than ours for sufficiently large models and that is in fact closer to ε_n^2 than ε_n, where ε_n is defined in Theorem 3.3. But the mean of the true excess risk has to be compared to the mean of the empirical excess risk and it is remarkable that in Proposition 12 of [10] where such a result is given in a way that seems very sharp, there is a term lower bounded by

$$\left(n \times \inf_{I \in \mathcal{P}} P^X\left(I\right)\right)^{-1/4} \propto \left(\frac{D}{n}\right)^{1/4} \ ,$$

due to the lower regularity assumption on the partition. This allows us to conjecture that up to a logarithmic factor, the term proportional to $\left(\frac{D\ln n}{n}\right)^{1/4}$ appearing in ε_n and also in the deviations of the true excess risk in Theorem 3.1 is not improvable in general, and that the empirical excess risk concentrates better around its mean than the true excess risk in general. We can conclude that the bounds given in Proposition 10, 11 and 12 of [10] are better than ours, apart for the bias term involved in concentration inequalities of Proposition 10, but this term could be removed as explained above. Furthermore, concentration inequalities for the empirical excess risk are significantly better than ours for large models.

Arlot and Massart [10] also propose generalizations in the case of unbounded noise and when the noise level vanishes. The unbounded case seems to be beyond the reach of our strategy, due to our repeated use of Bousquet and Klein-Rio's inequalities along the proofs. However, we recover the case of vanishing noise level for histogram models, when the partition is upper regular with respect to the law P^X, a condition also needed in [10] in this case. Indeed, we have noticed in Section 3.3.3 that assumption (**H2**) can be weaken by (**H2bis**) where we assume that

$$\mathcal{K}_{1,M} \geq A_{\min} > 0$$

for some positive constant $A_{\min}$. So, if we assume the upper regularity of the partition $\mathcal{P}$ with respect to P^X, that is

$$|\mathcal{P}| \sup_{I \in \mathcal{P}} P^X(I) \leq c^+_{M,P} < +\infty \tag{3.44}$$

for a positive constant $c^+_{M,P}$, we then have from identity (3.30)

$$\mathcal{K}^2_{1,M} \geq 4\mathbb{E}\left[\sigma^2(X)\,\Psi^2_M(X)\right] \ ,$$

where from identity (3.29), we have in the case of histograms,

$$\Psi^2_M(x) = \frac{1}{|\mathcal{P}|} \sum_{I \in \mathcal{P}} \frac{\mathbf{1}_{x \in I}}{P^X(I)} \ , \ \text{for all } x \in \mathcal{X} \ .$$

Now from inequality (3.44) we have

$$\Psi^2_M(x) \geq \left(c^+_{M,P}\right)^{-1} > 0 \ , \ \text{for all } x \in \mathcal{X} \ ,$$

and so $A_{\min} = 2\left(c^+_{M,P}\right)^{-1/2} \|\sigma\|_2 > 0$ is convenient in (**H2bis**).

3.5 The case of piecewise polynomials

In this Section, we generalize the results given in Section 3.4 for models of piecewise constant functions to models of piecewise polynomials uniformly bounded in their degree.

3.5.1 Existence of a localized basis

The following lemma states the existence of a localized orthonormal basis in $(M, \|\cdot\|_2)$ where M is a model of piecewise polynomials and $\mathcal{X} = [0,1]$ is the unit interval.

Lemma 3.4 *Let* Leb *denote the Lebesgue measure on* $[0,1]$. *Let assume that* $\mathcal{X} = [0,1]$ *and that* P^X *has a density* f *with respect to* Leb *satisfying, for a positive constant* $c_{\min}$,

$$f(x) \geq c_{\min} > 0, \quad x \in [0,1] \ .$$

Consider a linear model M *of piecewise polynomials on* $[0,1]$ *with degree* r *or smaller, defined on a finite partition* $\mathcal{P}$ *made of intervals. Then there exists an orthonormal basis* $\{\varphi_{I,j},\ I \in \mathcal{P},\ j \in \{0,...,r\}\}$ *of* $(M, \|\cdot\|_2)$ *such that,*

$$\text{for all } j \in \{0,...,r\} \qquad \varphi_{I,j} \text{ is supported by the element } I \text{ of } \mathcal{P},$$

and a constant $L_{r,c_{\min}}$ *depending only on* $r, c_{\min}$ *exists, satisfying for all* $I \in \mathcal{P}$,

$$\max_{j \in \{0,...,r\}} \|\varphi_{I,j}\|_\infty \leq L_{r,c_{\min}} \frac{1}{\sqrt{\text{Leb}(I)}} \ . \tag{3.45}$$

As a consequence, if it holds

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \mathrm{Leb}(I)} \geq c_{M,\mathrm{Leb}} \tag{3.46}$$

a constant $L_{r,c_{\min},c_{M,\mathrm{Leb}}}$ *depending only on* $r, c_{\min}$ *and* $c_{M,\mathrm{Leb}}$ *exists, such that for all* $\beta = (\beta_{I,j})_{I \in \mathcal{P}, j \in \{0,\ldots,r\}} \in \mathbb{R}^D$,

$$\left\| \sum_{I,j} \beta_{I,j} \varphi_{I,j} \right\|_\infty \leq L_{r,c_{\min},c_{M,\mathrm{Leb}}} \sqrt{D} \, |\beta|_\infty \tag{3.47}$$

where $D = (r+1)|\mathcal{P}|$ *is the dimension of* M.

Lemma 3.4 states that if $\mathcal{X} = [0,1]$ is the unit interval and P^X has a density with respect to the Lebesgue measure Leb on $\mathcal{X}$ uniformly bounded away form zero, then there exists an orthonormal basis in $L_2\left(P^X\right)$ of piecewise polynomials where the sup-norm of its elements are suitably controlled by (3.45). Moreover, if we assume the lower regularity of the partition with respect to Leb then the orthonormal basis is localized, where the constant of localization in (3.47) depend on the maximal degree r. We notice that in the case of piecewise constant functions we do not need to assume the existence of a density for P^X or to restrict ourselves to the unit interval. The proof of Lemma 3.4 can be found in Section 3.6.2.

3.5.2 Rates of convergence in sup-norm

The following lemma allows to derive property (**H5**) for piecewise polynomials.

Lemma 3.5 *Assume that Inequality (3.11) holds, that is a positive constant* A *exists such that* $|Y| \leq A$ *a.s. Denote by* Leb *the Lebesgue measure on* $[0,1]$. *Assume that* $\mathcal{X} = [0,1]$ *and that* P^X *has a density* f *with respect to* Leb, *satisfying for positive constants* $c_{\min}$ *and* $c_{\max}$,

$$0 < c_{\min} \leq f(x) \leq c_{\max} < +\infty, \ x \in [0,1] \ . \tag{3.48}$$

Consider a linear model M *of piecewise polynomials on* $[0,1]$ *with degree less than* r, *defined on a finite partition* $\mathcal{P}$ *made of intervals, that satisfies for some finite positive constants* $c_{M,\mathrm{Leb}}$

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \mathrm{Leb}(I)} \geq c_{M,\mathrm{Leb}} > 0 \ . \tag{3.49}$$

Assume moreover that $D \leq A_+ n (\ln n)^{-2}$ *for a positive finite constant* A_+. *Then, for any* $\alpha > 0$, *there exists an event of probability at least* $1 - n^{-\alpha}$ *such that* s_n *exists, is unique on this event and it holds, for all* $n \geq n_0(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \alpha)$,

$$\|s_n - s_M\|_\infty \leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\alpha} \sqrt{\frac{D \ln n}{n}} \ . \tag{3.50}$$

In Lemma 3.3 we thus obtain the convergence in sup-norm of the M-estimator s_n towards the linear projection s_M at the rate $\sqrt{\frac{D \ln n}{n}}$. It is worth noticing that for a model of piecewise polynomials satisfying the assumptions of Lemma 3.3, if we set

$$A_{cons} = L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\alpha} \sqrt{A_+} \quad , \quad R_{n,D,\alpha} = L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\alpha} \sqrt{\frac{D \ln n}{n}} \ ,$$
$$n_1 = n_0(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \alpha)$$

then Assumption **(H5)** is satisfied. To derive Inequality (3.38), we need to assume that the response variable Y is almost surely bounded, we give the conditions to ensure that the model is provided with a localized basis and also we assume that the density of P^X with respect to the Lebesgue measure on the unit interval is uniformly bounded from above. The proof of Lemma 3.5 can be found in Section 3.6.2.

3.5.3 Bounds for the excess risks

The forthcoming result is a straightforward application of Lemmas 3.4, 3.5 and Theorems 3.1, 3.2.

Theorem 3.4 *Denote by* Leb *the Lebesgue measure on* $[0,1]$ *and fix some positive finite constant* α. *Assume that* $\mathcal{X}=[0,1]$ *and that* P^X *has a density* f *with respect to* Leb *satisfying, for some positive finite constants* $c_{\min}$ *and* $c_{\max}$,

$$0 < c_{\min} \leq f(x) \leq c_{\max} < +\infty,\ x \in [0,1]\ . \tag{3.51}$$

Consider a linear model M *of piecewise polynomials on* $[0,1]$ *with degree less than* r, *defined on a finite partition* $\mathcal{P}$ *made of intervals, that satisfy for a finite constant* $c_{M,\mathrm{Leb}}$,

$$\sqrt{|\mathcal{P}| \inf_{I\in\mathcal{P}} \mathrm{Leb}(I)} \geq c_{M,\mathrm{Leb}} > 0\ . \tag{3.52}$$

*Assume that (**H1**) and (**H2**) hold. Then, if there exist some positive finite constants* A_- *and* A_+ *such that*

$$A_- (\ln n)^2 \leq D \leq A_+ \frac{n}{(\ln n)^2}\ ,$$

then there exists a positive finite constant A_0, *depending on* $\alpha, A, \sigma_{\min}, A_-, A_+, r, c_{M,\mathrm{Leb}}, c_{\min}$ *and* $c_{\max}$ *such that, by setting*

$$\varepsilon_n = A_0 \max\left\{ \left(\frac{\ln n}{D}\right)^{1/4}, \left(\frac{D\ln n}{n}\right)^{1/4} \right\}$$

we have, for all $n \geq n_0(A_-, A_+, A, r, \sigma_{\min}, c_{M,\mathrm{Leb}}, c_{\min}, c_{\max}, \alpha)$,

$$\mathbb{P}\left[(1+\varepsilon_n)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \geq P(Ks_n - Ks_M) \geq (1-\varepsilon_n)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \right] \geq 1 - 10n^{-\alpha}$$

and

$$\mathbb{P}\left[\left(1+\varepsilon_n^2\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \geq P_n(Ks_M - Ks_n) \geq \left(1-\varepsilon_n^2\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-\alpha}\ .$$

*Moreover, if (3.51) and (3.52) hold together with (**H1**) and if we assume that*

$$1 \leq D \leq A_+ \frac{n}{(\ln n)^2}\ ,$$

then a positive constant A_u *exists, only depending on* $A_+, A, r, c_{M,\mathrm{Leb}}, c_{\min}$ *and* α, *such that for all* $n \geq n_0(A_+, A, r, c_{\min}, c_{\max}, c_{M,\mathrm{Leb}}, \alpha)$,

$$\mathbb{P}\left[P(Ks_n - Ks_M) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3n^{-\alpha}$$

and

$$\mathbb{P}\left[P_n(Ks_M - Ks_n) \geq A_u \frac{D \vee \ln n}{n} \right] \leq 3n^{-\alpha}\ .$$

We derive in Theorem 3.4 optimal upper and lower bounds for the excess risk and its empirical counterpart in the case of models of piecewise polynomials uniformly bounded in their degree with reasonable dimension. We give also upper bounds for models of possibly small dimension, without assumption (**H2**). Notice that we need stronger assumptions than in the case of histograms. Namely, we require the existence of a density uniformly bounded from above and from below for the unknown law P^X, with respect to the Lebesgue measure on the unit interval.

However we recover the bounds of Theorem 3.3 yet with different constants, since by Lemma 3.5 we still have $R_{n,D,\alpha} \propto \sqrt{\frac{D \ln n}{n}}$ as in the histogram case.
Moreover, as in the case of histograms, assumption (3.12) which is part of (**H1**) is a straightforward consequence of (3.11). Indeed, we easily see that the projection s_M of the regression function s_* onto the model of piecewise polynomials with respect to $\mathcal{P}$ can be written

$$s_M = \sum_{(I,j)\in\mathcal{P}\times\{0,\dots,r\}} P\left(Y\varphi_{I,j}\right)\varphi_{I,j}$$

where $\varphi_{I,j}$ is the orthonormal basis given in Lemma 3.4. It is then easy to show, using (3.45) of Lemma 3.4 and (3.11), that $\|s_M\|_\infty \leq L_{A,r,c_{\min},c_{\max}}$.
Again, we can consider vanishing noise at the prize to ask that the partition is upper regular with respect to Leb. By (**H2bis**) of Section 4.3.3 if we show that

$$\mathcal{K}_{1,M} \geq A_{\min} > 0$$

for a positive constant $A_{\min}$ instead of (**H2**), then the conclusions of Theorem 3.4 still hold. Now, from identity (3.30) we have

$$\mathcal{K}_{1,M}^2 \geq 4\mathbb{E}\left[\sigma^2(X)\,\Psi_M^2(X)\right]$$

where from identity (3.29), it holds in the case of piecewise polynomials, for all $x \in \mathcal{X}$,

$$\Psi_M^2(x) = \frac{1}{(r+1)\,|\mathcal{P}|}\sum_{(I,j)\in\mathcal{P}\times\{0,\dots,r\}}\varphi_{I,j}^2 \geq \frac{1}{(r+1)\,|\mathcal{P}|}\sum_{I\in\mathcal{P}}\frac{\mathbf{1}_{x\in I}}{P^X(I)}\ . \tag{3.53}$$

Furthermore, if we ask that

$$|\mathcal{P}|\sup_{I\in\mathcal{P}}\mathrm{Leb}(I) \leq c_{M,P}^+ < +\infty \tag{3.54}$$

for a positive constant $c_{M,P}^+$, then by using (3.51), (3.53) and (3.54), we obtain for all $x \in \mathcal{X}$,

$$\Psi_M^2(x) \geq \Big(c_{\max}\times c_{M,P}^+\times(r+1)\Big)^{-1} > 0\ ,$$

and so $A_{\min} = 2\Big(c_{\max}\times c_{M,P}^+\times(r+1)\Big)^{-1/2}\sqrt{\mathbb{E}\left[\sigma^2(X)\right]} > 0$ is convenient in (**H2bis**).

3.6 Proofs

We begin with the simpler proofs of Sections 3.4 and 3.5, in Sections 3.6.1 and 3.6.2 respectively. The proofs of Theorems 3.1 and 3.2 of Section 3.3.2 can be found in Section 3.6.3.

3.6.1 Proofs of Section 3.4

Proof of Lemma 3.2. It suffices to observe that

$$\begin{aligned}\left\|\sum_{I\in\mathcal{P}}\beta_I\varphi_I\right\|_\infty &\leq |\beta|_\infty \sup_{I\in\mathcal{P}}\|\varphi_I\|_\infty \\ &= |\beta|_\infty \sup_{I\in\mathcal{P}}\left(P^X(I)\right)^{-1/2} \\ &\leq c_{M,P}^{-1}\sqrt{D}\,|\beta|_\infty\ .\end{aligned}$$

■

We now intend to prove (3.38) under the assumptions of Lemma 3.3.

Proof of Lemma 3.3. Along the proof, we denote by misuse of notation, for any $I \in \mathcal{P}$,

$$P(I) := P(I \times \mathbb{R}) = P^X(I) \text{ and } P_n(I) := P_n(I \times \mathbb{R}) \ .$$

Let $\alpha > 0$ be fixed and let $\beta > 0$ to be chosen later. We first show that, since we have $D \leq A_+ n (\ln n)^{-2}$, it holds with large probability and for all n sufficiently large,

$$\inf_{I \in \mathcal{P}} P_n(I) > 0 \ .$$

Since

$$\|\mathbf{1}_I\|_\infty \leq 1 \quad \text{and} \quad \mathbb{E}\left[\mathbf{1}_I^2\right] = P(I)$$

we get by Bernstein's inequality (7.46), for any $x > 0$ and $I \in \mathcal{P}$,

$$\mathbb{P}\left[|(P_n - P)(I)| \geq \sqrt{\frac{2P(I)\,x}{n}} + \frac{x}{3n}\right] \leq 2\exp(-x) \ . \tag{3.55}$$

Further note that by (3.37), $D \geq c_{M,P}^2 P(I)^{-1} > 0$ for any $I \in \mathcal{P}$, and thus by taking $x = \beta \ln n$, we easily deduce from inequality (3.55) that there exists a positive constant $L^{(1)}_{\beta, c_{M,P}}$ only depending on $c_{M,P}$ and β such that, for any $I \in \mathcal{P}$,

$$\mathbb{P}\left[\frac{|(P_n - P)(I)|}{P(I)} \geq L^{(1)}_{\beta, c_{M,P}} \sqrt{\frac{D \ln n}{n}}\right] \leq 2n^{-\beta} \ . \tag{3.56}$$

Now, as $D \leq A_+ n (\ln n)^{-2}$ for some positive constant A_+, a positive integer $n_0(\beta, c_{M,P}, A_+)$ exists such that

$$L^{(1)}_{\beta, c_{M,P}} \sqrt{\frac{D \ln n}{n}} \leq \frac{1}{2}, \text{ for all } n \geq n_0(\beta, c_{M,P}, A_+) \ . \tag{3.57}$$

Therefore we get, for all $n \geq n_0(\beta, c_{M,P}, A_+)$,

$$\begin{aligned}
&\mathbb{P}[\forall I \in \mathcal{P},\ P_n(I) > 0] \\
&\geq \mathbb{P}\left[\forall I \in \mathcal{P},\ \frac{P(I)}{2} > |(P_n - P)(I)|\right] \\
&\geq \mathbb{P}\left[\forall I \in \mathcal{P},\ \frac{|(P_n - P)(I)|}{P(I)} < L^{(1)}_{\beta, c_{M,P}} \sqrt{\frac{D \ln n}{n}}\right] \quad \text{by (3.57)} \\
&\geq 1 - 2Dn^{-\beta} \ .
\end{aligned}$$

Introduce the event

$$\Omega_+ = \{\forall I \in \mathcal{P},\ P_n(I) > 0\} \ .$$

We have shown that

$$\mathbb{P}[\Omega_+] \geq 1 - 2Dn^{-\beta} \ . \tag{3.58}$$

Moreover, on the event Ω_+, the least-squares estimator s_n exists, is unique and it holds

$$s_n = \sum_{I \in \mathcal{P}} \frac{P_n(y\mathbf{1}_{x \in I})}{P_n(I)} \mathbf{1}_I \ .$$

We also have

$$s_M = \sum_{I \in \mathcal{P}} \frac{P(y\mathbf{1}_{x \in I})}{P(I)} \mathbf{1}_I \ .$$

Hence it holds on Ω_+,

$$\begin{aligned}\|s_n - s_M\|_\infty &= \sup_{I\in\mathcal{P}}\left|\frac{P_n\left(y\mathbf{1}_{x\in I}\right)}{P_n\left(I\right)} - \frac{P\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)}\right| \\ &= \sup_{I\in\mathcal{P}}\left|\frac{P_n\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)\left(1+\frac{(P_n-P)(I)}{P(I)}\right)} - \frac{P\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)}\right| \\ &\leq \sup_{I\in\mathcal{P}}\left|\frac{\left(P_n-P\right)\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)\left(1+\frac{(P_n-P)(I)}{P(I)}\right)}\right| \\ &\quad + \sup_{I\in\mathcal{P}}\left|\frac{P\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)}\right| \times \sup_{I\in\mathcal{P}}\left|1 - \frac{1}{1+\frac{(P_n-P)(I)}{P(I)}}\right| . \end{aligned} \tag{3.59}$$

Moreover, by Bernstein's inequality (7.46), as

$$\|y\mathbf{1}_{x\in I}\|_\infty \leq A \qquad \text{and} \qquad \mathbb{E}\left[\left(Y\mathbf{1}_{X\in I}\right)^2\right] \leq A^2 P\left(I\right)$$

we get for all $I \in \mathcal{P}$,

$$\mathbb{P}\left[\left|\left(P_n - P\right)\left(y\mathbf{1}_{x\in I}\right)\right| \geq \sqrt{\frac{2A^2P\left(I\right)x}{n}} + \frac{Ax}{3n}\right] \leq 2\exp\left(-x\right) .$$

By putting $x = \beta \ln n$ in the latter inequality and using the fact that $D \geq c_{M,P}^2 P\left(I\right)^{-1}$ it follows that there exists a positive constant $L^{(2)}_{A,c_{M,P},\beta}$ only depending on A, $c_{M,P}$ and β such that

$$\mathbb{P}\left[\frac{\left|\left(P_n - P\right)\left(y\mathbf{1}_{x\in I}\right)\right|}{P\left(I\right)} \geq L^{(2)}_{A,c_{M,P},\beta}\sqrt{\frac{D\ln n}{n}}\right] \leq 2n^{-\beta} . \tag{3.60}$$

Now define

$$\Omega_{1,2} = \bigcap_{I\in\mathcal{P}}\left\{\left\{\frac{\left|\left(P_n - P\right)\left(I\right)\right|}{P\left(I\right)} < L^{(1)}_{\beta,c_{M,P}}\sqrt{\frac{D\ln n}{n}}\right\}\bigcap\left\{\frac{\left|\left(P_n - P\right)\left(y\mathbf{1}_{x\in I}\right)\right|}{P\left(I\right)} < L^{(2)}_{A,c_{M,P},\beta}\sqrt{\frac{D\ln n}{n}}\right\}\right\} .$$

Clearly, since $D \leq n$ we have, by (3.56) and (3.60),

$$\mathbb{P}\left[\Omega_{1,2}^c\right] \leq 4n^{-\beta+1} . \tag{3.61}$$

Moreover, for all $n \geq n_0\left(\beta, c_{M,P}, A_+\right)$, we get by (3.57) that

$$\frac{\left|\left(P_n - P\right)\left(I\right)\right|}{P\left(I\right)} < \frac{1}{2}$$

on the event $\Omega_{1,2}$, and so, for all $n \geq n_0\left(\beta, c_{M,P}, A_+\right)$, $\Omega_{1,2} \subset \Omega_+$. Hence, we get that

$$\begin{aligned}&\sup_{I\in\mathcal{P}}\left|\frac{\left(P_n - P\right)\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)\left(1+\frac{(P_n-P)(I)}{P(I)}\right)}\right| + \sup_{I\in\mathcal{P}}\left|\frac{P\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)}\right| \times \sup_{I\in\mathcal{P}}\left|1 - \frac{1}{1+\frac{(P_n-P)(I)}{P(I)}}\right| \\ &\leq 2\sup_{I\in\mathcal{P}}\left|\frac{\left(P_n - P\right)\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)}\right| + 2\sup_{I\in\mathcal{P}}\left|\frac{P\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)}\right| \times \sup_{I\in\mathcal{P}}\left|\frac{\left(P_n - P\right)\left(I\right)}{P\left(I\right)}\right| \\ &\leq 2L^{(2)}_{A,c_{M,P},\beta}\sqrt{\frac{D\ln n}{n}} + 2L^{(1)}_{\beta,c_{M,P}}\sqrt{\frac{D\ln n}{n}} \times \sup_{I\in\mathcal{P}}\left|\frac{P\left(y\mathbf{1}_{x\in I}\right)}{P\left(I\right)}\right| . \end{aligned} \tag{3.62}$$

Finally we have, for any $I \in \mathcal{P}$,

$$|P(y\mathbf{1}_{x\in I})| \leq P(|y|\,\mathbf{1}_{x\in I}) \leq AP(I)\,, \tag{3.63}$$

so by (3.59), (3.62) and (3.63) we finally get, on the event $\Omega_{1,2}$ and for all $n \geq n_0(\beta, c_{M,P}, A_+)$,

$$\|s_n - s_M\|_\infty \leq \left(2L^{(2)}_{A,c_{M,P},\beta} + 2AL^{(1)}_{\beta,c_{M,P}}\right)\sqrt{\frac{D\ln n}{n}}\,.$$

Taking $\beta = \alpha + 3$, we get by (3.61) for all $n \geq 2$, $\mathbb{P}\left[\Omega^c_{1,2}\right] \leq n^{-\alpha}$ which implies (3.38).

■

3.6.2 Proofs of Section 3.5

Under the assumptions of Lemma 3.4, we intend to establish (3.47).

Proof of Lemma 3.4. Let I be any interval of $[0,1]$ and w a positive measurable function on I. Denote by $L_2(I, \text{Leb})$ the space of square integrable functions on I with respect to the Lebesgue measure Leb and set

$$L_2(I,w) = \left\{g : I \longrightarrow \mathbb{R}\ ;\ g\sqrt{w} \in L_2(I, \text{Leb})\right\}\ .$$

This space is equipped with the natural inner product

$$\langle g,h\rangle_{I,w} = \int_{x\in I} g(x)\,h(x)\,w(x)\,dx\ .$$

Write $\|.\|_{I,w}$ its associated norm.
Now, consider an interval I of $\mathcal{P}$ with bounds a and b, $a < b$. Also denote by $f_{|I} : x \in I \longmapsto f(x)$ the restriction of the density f to the interval I. We readily have for $g, h \in L_2\left(I, f_{|I}\right)$,

$$\begin{aligned}&\int_{x\in I} g(x)\,h(x)\,f_{|I}(x)\,\frac{dx}{\text{Leb}(I)}\\ &= \int_{y\in[0,1]} g((b-a)\,y+a)\,h((b-a)\,y+a)\,f_{|I}((b-a)\,y+a)\,dy\ .\end{aligned} \tag{3.64}$$

Define the function f^I from $[0,1]$ to $\mathbb{R}_+$ by

$$f^I(y) = f_{|I}((b-a)\,y+a)\,,\quad y\in[0,1]\ .$$

If $(p_{I,0}, p_{I,1}, ...p_{I,r})$ is an orthonormal family of polynomials in $L_2\left([0,1], f^I\right)$ then by setting, for all $x \in I$, $j \in \{0,...,r\}$,

$$\tilde{\varphi}_{I,j}(x) = p_{I,j}\left(\frac{x-a}{b-a}\right)\frac{1}{\sqrt{\text{Leb}(I)}}\ ,$$

we deduce from equality (3.64) that $(\tilde{\varphi}_{I,j})_{j=0}^r$ is an orthonormal family of polynomials in $L_2\left(I, f_{|I}\right)$ such that $\deg(\tilde{\varphi}_{I,j}) = \deg(p_{I,j})$.
Now, it is a classical fact of orthogonal polynomials theory (see for example Theorems 1.11 and 1.12 of [33]) that there exists a unique family $(q_{I,0}, q_{I,1}, ...q_{I,r})$ of orthogonal polynomials

on $[0,1]$ such that $\deg(q_{I,j}) = j$ and the coefficient of the highest monomial x^j of $q_{I,j}$ is equal to 1. Moreover, each $q_{I,j}$ has j distinct real roots belonging to $]0,1[$. Thus, we can write

$$q_{I,j}(x) = \prod_{k=1}^{j} \left(x - \alpha_{I,j}^k\right), \quad \alpha_{I,j}^k \in]0,1[\text{ and } \alpha_{I,j}^k \neq \alpha_{I,j}^l \text{ for } k \neq l \ . \tag{3.65}$$

Clearly, $\|q_{I,j}\|_\infty \leq 1$. Moreover,

$$\begin{aligned} \|q_{I,j}\|_{[0,1],f^I}^2 &= \int_{[0,1]} (q_{I,j})^2 f^I dx \\ &\geq c_{\min} \int_{[0,1]} (q_{I,j})^2 dx \ . \end{aligned}$$

Now we set $B(\alpha, r) =]\alpha - r, \alpha + r[$ for $\alpha \in \mathbb{R}$, so that by (3.65) we get

$$\forall x \in [0,1] \setminus \cup_{k=1}^j B\left(\alpha_{I,j}^k, (4j)^{-1}\right), \quad |q_{I,j}(x)| \geq (4j)^{-j} \ ,$$

and

$$\text{Leb}\left([0,1] \setminus \cup_{k=1}^j B\left(\alpha_{I,j}^k, (4j)^{-1}\right)\right) \geq \frac{1}{2} \ .$$

Therefore,

$$\begin{aligned} \|q_{I,j}\|_{[0,1],f^I}^2 &\geq c_{\min} \int_{[0,1]} (q_{I,j})^2 dx \\ &\geq c_{\min} \int_{[0,1]\setminus\cup_{k=1}^j B\left(\alpha_{I,j}^k,(4j)^{-1}\right)} (q_{I,j})^2 dx \\ &\geq \frac{c_{\min}}{2} (4j)^{-2j} \ . \end{aligned}$$

Finally, introduce $p_{I,j} = \|q_{I,j}\|_{[0,1],f^I}^{-1} q_{I,j}$ and denote by $\varphi_{I,j}$ its associated orthonormal family of $L_2\left(I, f_{|I}\right)$. Then, by considering the extension $\varphi_{I,j}$ of $\tilde{\varphi}_{I,j}$ to $[0,1]$ by adding null values, it is readily checked that the family

$$\{\varphi_{I,j}, \ I \in \mathcal{P}, \ j \in \{0, ..., r\}\}$$

is an orthonormal basis of $(M, \|\cdot\|_2)$. In addition,

$$\begin{aligned} \|\varphi_{I,j}\|_\infty &= \|\tilde{\varphi}_{I,j}\|_\infty \\ &= \|q_{I,j}\|_{[0,1],f^I}^{-1} \|q_{I,j}\|_\infty \text{Leb}(I)^{-1/2} \\ &\leq \sqrt{2} c_{\min}^{-1/2} (4r)^r \text{Leb}(I)^{-1/2} \qquad (3.66) \\ &\leq \sqrt{2} c_{M,\text{Leb}}^{-1} c_{\min}^{-1/2} (4r)^r (r+1)^{-1/2} \sqrt{D} \qquad (3.67) \end{aligned}$$

where in the last inequality we used the fact that

$$\sqrt{|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb}(I)} \geq c_{M,\text{Leb}} \ \text{ and } \ D = (r+1)|\mathcal{P}| \ .$$

For all $j \in \{0, ..., r\}$, $\varphi_{I,j}$ is supported by the element I of $\mathcal{P}$, hence we deduce from (3.66) that the orthonormal basis $\{\varphi_{I,j}, \ I \in \mathcal{P}, \ j \in \{0, ..., r\}\}$ of $(M, \|\cdot\|_2)$ satisfies (3.45) with

$$L_{r,c_{\min}} = \sqrt{2} c_{\min}^{-1/2} (4r)^r \ .$$

To conclude, observe that

$$\begin{aligned}\left\|\sum_{I,j}\beta_{I,j}\varphi_{I,j}\right\|_\infty &= \max_{I\in\mathcal{P}}\left\{\left\|\sum_{j=0}^{r}\beta_{I,j}\varphi_{I,j}\right\|_\infty\right\}\\ &\leq |\beta|_\infty \max_{I\in\mathcal{P}}\left\{\sum_{j=0}^{r}\|\varphi_{I,j}\|_\infty\right\}\\ &\leq (r+1)\,|\beta|_\infty \max_{I\in\mathcal{P}}\max_{j\in\{0,\dots,r\}}\left\{\|\varphi_{I,j}\|_\infty\right\}\end{aligned}$$

and thus, by plugging (3.67) into the right-hand side of the last inequality, we finally obtain that the value

$$L_{r,c_{\min},c_{M,\mathrm{Leb}}} = \sqrt{2}c_{M,\mathrm{Leb}}^{-1}c_{\min}^{-1/2}\,(4r)^r\,(r+1)^{1/2}$$

gives the desired bound (3.47). ■

We now turn to the proof of (3.50) under the assumptions of Lemma 3.5. The proof is based on concentration inequalities recalled in Section 7.4.1 of Chapter 7 and on inequality (3.45) of Lemma 3.4, that allows us to control the sup-norm of elements of an orthonormal basis for a model of piecewise polynomials.

Proof of Lemma 3.5. Let $\alpha > 0$ be fixed and $\gamma > 0$ to be chosen later. The partition $\mathcal{P}$ associated to M will be denoted by

$$\mathcal{P} = \{I_0, ..., I_{m-1}\}\ ,$$

so that $|\mathcal{P}| = m$ and $D = (r+1)\,m$ where D is the dimension of the model M. By (3.45) of Lemma 3.4 there exists an orthonormal basis $\{\varphi_{I_k,j};\ k \in \{0,...,m-1\}\,,\ j \in \{0,...,r\}\}$ of $\left(M, L_2\left(P^X\right)\right)$ such that,

$$\varphi_{I_k,j} \text{ is supported by the element } I_k \text{ of } \mathcal{P}, \quad \text{for all } j \in \{0,...,r\}$$

and a constant $L_{r,c_{\min}}$ depending only on $r, c_{\min}$ and satisfying

$$\max_{j\in\{0,\dots,r\}}\|\varphi_{I_k,j}\|_\infty \leq L_{r,c_{\min}}\frac{1}{\sqrt{\mathrm{Leb}\,(I_k)}},\ \text{ for all } k \in \{0,...,m-1\}\,. \tag{3.68}$$

In order to avoid cumbersome notation, we define a total ordering $\preceq$ on the set

$$\mathcal{I} = \{(I_k, j)\,; k \in \{0,...,m-1\}\,,\ j \in \{0,...,r\}\}\ ,$$

as follows. Let $\prec$ be a binary relation on $\mathcal{I}\times\mathcal{I}$ such that

$$(I_k, j) \prec (I_l, i) \quad \text{if} \quad (k < l \quad \text{or} \quad (k = l \quad \text{and} \quad j < i))\,,$$

and consider the total ordering $\preceq$ defined to be

$$(I_k, j) \preceq (I_l, i) \quad \text{if} \quad ((I_k, j) = (I_l, i) \quad \text{or} \quad (I_k, j) \prec (I_l, i))\ .$$

So, from the definition of $\preceq$, the vector $\beta = (\beta_{I_k,j})_{(I_k,j)\in\mathcal{I}} \in \mathbb{R}^D$ has coordinate $\beta_{I_k,j}$ at position $(r+1)\,k + j + 1$ and when the matrix

$$A = \left(A_{(I_k,j),(I_l,i)}\right)_{(I_k,j),(I_l,i)\in\mathcal{I}\times\mathcal{I}} \in \mathbb{R}^{D\times D}\ ,$$

has coefficient $A_{(I_k,j),(I_l,i)}$ at line $(r+1)\,k+j+1$ and column $(r+1)\,l+i+1$.
Now, for some $s=\sum_{(I_k,j)\in\mathcal{I}}\beta_{I_k,j}\varphi_{I_k,j}\in M$, we have

$$\begin{aligned}P_n\left(K\left(s\right)\right) &= P_n\left[\left(y-\left(\sum_{(I_k,j)\in\mathcal{I}}\beta_{I_k,j}\varphi_{I_k,j}\left(x\right)\right)\right)^2\right]\\ &= P_n y^2-2\sum_{(I_k,j)\in\mathcal{I}}\beta_{I_k,j}P_n\left(y\varphi_{I_k,j}\left(x\right)\right)+\sum_{(I_k,j),(I_l,i)\in\mathcal{I}\times\mathcal{I}}\beta_{I_k,j}\beta_{I_l,i}P_n\left(\varphi_{I_k,j}\varphi_{I_l,i}\right)\ .\end{aligned}$$

Hence, by taking the derivative with respect to $\beta_{I_k,j}$ in the last quantity,

$$\begin{aligned}&\frac{1}{2}\frac{\partial}{\partial\beta_{I_k,j}}P_n\left[\left(y-\left(\sum_{(I_k,j)\in\mathcal{I}}\beta_{I_k,j}\varphi_{I_k,j}\left(x\right)\right)\right)^2\right]\\ &\quad=-P_n\left(y\varphi_{I_k,j}\left(x\right)\right)+\sum_{(I_l,i)\in\mathcal{I}}\beta_{I_l,i}P_n\left(\varphi_{I_k,j}\varphi_{I_l,i}\right)\ .\end{aligned}\tag{3.69}$$

We see that if $\beta^{(n)}=\left(\beta^{(n)}_{I_k,j}\right)_{(I_k,j)\in\mathcal{I}}\in\mathbb{R}^D$ is a critical point of

$$P_n\left[\left(y-\left(\sum_{(I_k,j)\in\mathcal{I}}\beta_{I_k,j}\varphi_{I_k,j}\left(x\right)\right)\right)^2\right]\ ,$$

it holds

$$\left(\frac{\partial}{\partial\beta_{I_k,j}}P_n\left[\left(y-\left(\sum_{(I_k,j)\in\mathcal{I}}\beta_{I_k,j}\varphi_{I_k,j}\left(x\right)\right)\right)^2\right]\right)\left(\beta^{(n)}\right)=0$$

and by combining (3.69) with the fact that

$$P\left(\varphi_{I_k,j}\right)^2=1\ ,\ \text{for all}\ \ (I_k,j)\in\mathcal{I}\quad\text{and}\quad P\left(\varphi_{I_k,j}\varphi_{I_l,i}\right)=0\ \ \text{if}\ \ (I_k,j)\neq(I_l,i)\ ,$$

we deduce that $\beta^{(n)}$ satisfies the following random linear system,

$$\left(I_D+L_{n,D}\right)\beta^{(n)}=X_{y,n}\tag{3.70}$$

where $X_{y,n}=\left(P_n\left(y\varphi_{I_k,j}\left(x\right)\right)\right)_{(I_k,j)\in\mathcal{I}}\in\mathbb{R}^D$, I_D is the identity matrix of dimension D and $L_{n,D}=\left(\left(L_{n,D}\right)_{(I_k,j),(I_l,i)}\right)_{(I_k,j),(I_l,i)\in\mathcal{I}\times\mathcal{I}}$ is a $D\times D$ matrix satisfying

$$\left(L_{n,D}\right)_{(I_k,j),(I_l,i)}=\left(P_n-P\right)\left(\varphi_{I_k,j}\varphi_{I_l,i}\right)\ .$$

Now, by inequality (3.82) in Lemma 3.6 below, one can find a positive integer $n_0\left(r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma\right)$ such that on an event Ω_n of probability at least $1-3Dn^{-\gamma}$, we have

$$\left\|L_{n,D}\right\|\leq\frac{1}{2}\ ,\tag{3.71}$$

where for a $D\times D$ matrix L, the operator norm $\|\cdot\|$ associated to the sup-norm on vectors is

$$\left\|L\right\|=\sup_{x\neq0}\frac{\left|Lx\right|_\infty}{\left|x\right|_\infty}\ .$$

Then we deduce from (3.71) that $\left(I_D+L_{n,D}\right)$ is a non-singular $D\times D$ matrix and, as a consequence, that the linear system (3.70) admits a unique solution $\beta^{(n)}$ on Ω_n for all $n_0\left(r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma\right)$.

Moreover, since $P_n\left(y-\left(\sum_{(I_k,j)\in\mathcal{I}}\beta_{I_k,j}\varphi_{I_k,j}(x)\right)\right)^2$ is a nonnegative quadratic functional with respect to $(\beta_{I_k,j})_{(I_k,j)\in\mathcal{I}}\in\mathbb{R}^D$ we can easily deduce that on Ω_n, $\beta^{(n)}$ achieves the unique minimum of $P_n\left(y-\left(\sum_{(I_k,j)\in\mathcal{I}}\beta_{I_k,j}\varphi_{I_k,j}(x)\right)\right)^2$ on $\mathbb{R}^D$. In other words,

$$s_n=\sum_{(I_k,j)\in\mathcal{I}}\beta^{(n)}_{I_k,j}\varphi_{I_k,j}$$

is the unique least-squares estimator on M, and by (3.70) it holds,

$$\beta^{(n)}_{I_k,j}\left(1+\sum_{(I_l,i)\in\mathcal{I}}(P_n-P)(\varphi_{I_k,j}\varphi_{I_l,i})\right)=P_n(y\varphi_{I_k,j}(x))\ \text{, for all }\ (I_k,j)\in\mathcal{I}. \tag{3.72}$$

Now, as $\varphi_{I_k,j}$ and $\varphi_{I_l,i}$ have disjoint supports when $k\neq l$, it holds $\varphi_{I_k,j}\varphi_{I_l,i}=0$ whenever $k\neq l$, and so equation (3.72) reduces to

$$\beta^{(n)}_{I_k,j}\times\left(1+\sum_{i=0}^{r}(P_n-P)(\varphi_{I_k,j}\varphi_{I_k,i})\right)=P_n(y\varphi_{I_k,j}(x))\ ,\ \text{for all }\ (I_k,j)\in\mathcal{I}\ . \tag{3.73}$$

Moreover, recalling that $s_M=\sum_{(I_k,j)\in\mathcal{I}}P(y\varphi_{I_k,j}(x))\varphi_{I_k,j}$, it holds

$$\begin{aligned}\|s_n-s_M\|_\infty&=\left\|\sum_{(I_k,j)\in\mathcal{I}}\left(\beta^{(n)}_{I_k,j}-P(y\varphi_{I_k,j}(x))\right)\varphi_{I_k,j}\right\|_\infty\\&\leq\max_{k\in\{0,\dots,m-1\}}\left\|\sum_{j=0}^{r}\left(\beta^{(n)}_{I_k,j}-P(y\varphi_{I_k,j}(x))\right)\varphi_{I_k,j}\right\|_\infty\\&\leq(r+1)\max_{k\in\{0,\dots,m-1\}}\left\{\left(\max_{j\in\{0,\dots,r\}}\left|\beta^{(n)}_{I_k,j}-P(y\varphi_{I_k,j}(x))\right|\right)\right.\\&\qquad\left.\times\max_{j\in\{0,\dots,r\}}\|\varphi_{I_k,j}\|_\infty\right\}\end{aligned} \tag{3.74}$$

where the first inequality comes from the fact that $\varphi_{I_k,j}$ and $\varphi_{I_l,i}$ have disjoint supports when $k\neq l$. We next turn to the control of the right-hand side of (3.74). Let the index (I_k,j) be fixed. By subtracting the quantity $(1+\sum_{i=0}^{r}(P_n-P)(\varphi_{I_k,j}\varphi_{I_k,i}))\times P(y\varphi_{I_k,j}(x))$ in each side of equation (3.73), we get

$$\begin{aligned}&\left(\beta^{(n)}_{I_k,j}-P(y\varphi_{I_k,j}(x))\right)\times\left(1+\sum_{i=0}^{r}(P_n-P)(\varphi_{I_k,j}\varphi_{I_k,i})\right)\\&=(P_n-P)(y\varphi_{I_k,j}(x))-\left(\sum_{i=0}^{r}(P_n-P)(\varphi_{I_k,j}\varphi_{I_k,i})\right)\times P(y\varphi_{I_k,j}(x))\ .\end{aligned} \tag{3.75}$$

Moreover, by Inequality (3.83) of Lemma 3.6, we have for all $n\geq n_0(r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma)$,

$$\sum_{i=0}^{r}|(P_n-P)(\varphi_{I_k,j}\varphi_{I_k,i})|\leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n\,\mathrm{Leb}(I_k)}}\leq\frac{1}{2} \tag{3.76}$$

on the event Ω_n. We thus deduce that

$$\left|\left(\beta^{(n)}_{I_k,j}-P(y\varphi_{I_k,j}(x))\right)\times\left(1+\sum_{i=0}^{r}(P_n-P)(\varphi_{I_k,j}\varphi_{I_k,i})\right)\right|\geq\frac{1}{2}\left|\beta^{(n)}_{I_k,j}-P(y\varphi_{I_k,j}(x))\right| \tag{3.77}$$

and

$$\left|\left(\sum_{i=0}^{r}\left(P_n-P\right)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right)\times P\left(y\varphi_{I_k,j}\left(x\right)\right)\right|\leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n\,\mathrm{Leb}\left(I_k\right)}}\times\left|P\left(y\varphi_{I_k,j}\left(x\right)\right)\right| \ . \tag{3.78}$$

Moreover, by (3.11), (3.48) and (3.68) we have

$$\begin{aligned}\left|P\left(y\varphi_{I_k,j}\left(x\right)\right)\right| &\leq A\left\|\varphi_{I_k,j}\right\|_\infty P\left(I_k\right)\\ &\leq Ac_{\max}\left\|\varphi_{I_k,j}\right\|_\infty \mathrm{Leb}\left(I_k\right)\\ &\leq Ac_{\max}L_{r,c_{\min}}\sqrt{\mathrm{Leb}\left(I_k\right)}\\ &\leq L_{A,r,c_{\min},c_{\max}}\sqrt{\mathrm{Leb}\left(I_k\right)} \ .\end{aligned} \tag{3.79}$$

Putting inequality (3.79) in (3.78) we obtain

$$\left|\left(\sum_{i=0}^{r}\left(P_n-P\right)\left(\varphi_{I_k,j}\varphi_{I_k,i}\right)\right)\times P\left(y\varphi_{I_k,j}\left(x\right)\right)\right|\leq L_{r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n}} \ . \tag{3.80}$$

Hence, using inequalities (3.77), (3.80) and inequality (3.84) of Lemma 3.6 in equation (3.75), we obtain that

$$\left|\beta_{I_k,j}^{(n)}-P\left(y\varphi_{I_k,j}\left(x\right)\right)\right|\leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n}}$$

on Ω_n. Since the constant $L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\gamma}$ does not depend on the index (I_k,j) we deduce by (3.68) that

$$\begin{aligned}&\left(\max_{j\in\{0,\dots,r\}}\left|\beta_{I_k,j}^{(n)}-P\left(y\varphi_{I_k,j}\left(x\right)\right)\right|\right)\times\max_{j\in\{0,\dots,r\}}\left\|\varphi_{I_k,j}\right\|_\infty\\ &\leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n}}\times\max_{j\in\{0,\dots,r\}}\left\|\varphi_{I_k,j}\right\|_\infty\\ &\leq L_{A,r,A_+,c_{\min},c_{\max},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n\,\mathrm{Leb}\left(I_k\right)}} \ .\end{aligned} \tag{3.81}$$

Finally, by using (3.49) and (3.81) in (3.74), we get for all $n\geq n_0\left(r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma\right)$, on the event Ω_n of probability at least $1-3Dn^{-\gamma}$,

$$\begin{aligned}\left\|s_n-s_M\right\|_\infty &\leq (r+1)\max_{k\in\{0,\dots,m-1\}}\left\{\left(\max_{j\in\{0,\dots,r\}}\left|\beta_{I_k,j}^{(n)}-P\left(y\varphi_{I_k,j}\left(x\right)\right)\right|\right)\times\max_{j\in\{0,\dots,r\}}\left\|\varphi_{I_k,j}\right\|_\infty\right\}\\ &\leq L_{A,r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n}}\max_{k\in\{0,\dots,m-1\}}\frac{1}{\sqrt{\mathrm{Leb}\left(I_k\right)}}\\ &\leq L_{A,r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{|\mathcal{P}|\ln n}{n}}\\ &\leq L_{A,r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{D\ln n}{n}} \ .\end{aligned}$$

To conclude, simply take $\gamma=\frac{\ln 3}{\ln 2}+\alpha+1$, so that it holds for $n\geq 2$, $\mathbb{P}\left[\Omega_n^c\right]\leq n^{-\alpha}$ which implies (3.50).
It remains to prove the following lemma that has been used all along the proof.

Lemma 3.6 *Recall that* $L_{n,D}=\left(\left(L_{n,D}\right)_{(I_k,j),(I_l,i)}\right)_{(I_k,j),(I_l,i)\in\mathcal{I}\times\mathcal{I}}$ *is a* $D\times D$ *matrix such that for all* $(k,l)\in\{0,\dots,m-1\}^2$, $(j,i)\in\{0,\dots,r\}^2$,

$$\left(L_{n,D}\right)_{(I_k,j),(I_l,i)}=\left(P_n-P\right)\left(\varphi_{I_k,j}\varphi_{I_l,i}\right) \ .$$

Also recall that for a $D \times D$ matrix L, the operator norm $\|\cdot\|$ associated to the sup-norm on the vectors is

$$\|L\| = \sup_{x \neq 0} \frac{|Lx|_\infty}{|x|_\infty} \ .$$

Then, under the assumptions of Lemma 3.5, a positive integer $n_0(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma)$ exists such that, for all $n \geq n_0(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma)$, the following inequalities hold on an event Ω_n of probability at least $1 - 3Dn^{-\gamma}$,

$$\|L_{n,D}\| \leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{D \ln n}{n}} \leq \frac{1}{2} \tag{3.82}$$

and for all $k \in \{0, ..., m-1\}$,

$$\max_{j \in \{0,...,r\}} \left\{ \sum_{i=0}^{r} \left| (P_n - P)(\varphi_{I_k,j} \varphi_{I_k,i}) \right| \right\} \leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{\ln n}{n \,\mathrm{Leb}(I_k)}} \leq \frac{1}{2} \ , \tag{3.83}$$

$$\max_{j \in \{0,...,r\}} \left| (P_n - P)(y \varphi_{I_k,j}(x)) \right| \leq L_{A,A_+,r,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{\ln n}{n}} \ . \tag{3.84}$$

Proof of Lemma 3.6. Let us begin with the proof of inequality (3.84). Let the index $(I_k, j) \in \mathcal{I}$ be fixed. By using Bernstein's inequality (7.46) and observing that, by (3.11),

$$\mathrm{Var}(y \varphi_{I_k,j}(x)) \leq P\left[(y \varphi_{I_k,j}(x))^2 \right] \leq \|Y\|_\infty^2 \leq A^2$$

and, by (3.11), (3.68) and (3.49),

$$\begin{aligned} \|Y \varphi_{I_k,j}(X)\|_\infty &\leq A \|\varphi_{I_k,j}(X)\|_\infty \\ &\leq A L_{r,c_{\min}} \frac{1}{\sqrt{\mathrm{Leb}(I_k)}} \\ &\leq L_{A,r,c_{\min},c_{M,\mathrm{Leb}}} \sqrt{|\mathcal{P}|} \\ &\leq L_{A,r,c_{\min},c_{M,\mathrm{Leb}}} \sqrt{D} \ , \end{aligned}$$

we get

$$\mathbb{P}\left[\left| (P_n - P)(y \varphi_{I_k,j}(x)) \right| \geq \sqrt{2A^2 \frac{x}{n}} + \frac{L_{A,r,c_{\min},c_{M,\mathrm{Leb}}} \sqrt{D}}{3n} x \right] \leq 2 \exp(-x) \ . \tag{3.85}$$

By taking $x = \gamma \ln n$ in inequality (3.85), we obtain that

$$\mathbb{P}\left[\left| (P_n - P)(y \varphi_{I_k,j}(x)) \right| \geq \sqrt{2A^2 \gamma \frac{\ln n}{n}} + \frac{L_{A,r,c_{\min},c_{M,\mathrm{Leb}}} \sqrt{D} \gamma \ln n}{3n} \right] \leq 2n^{-\gamma} \ . \tag{3.86}$$

Now, as $D \leq A_+ n (\ln n)^{-2}$, we deduce from (3.86) that for some well chosen positive constant $L_{A,A_+,r,c_{\min},c_{M,\mathrm{Leb}},\gamma}$, we have

$$\mathbb{P}\left[\left| (P_n - P)(y \varphi_{I_k,j}(x)) \right| \geq L_{A,A_+,r,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{\ln n}{n}} \right] \leq 2n^{-\gamma}$$

and by setting

$$\Omega_n^{(1)} = \bigcap_{(I_k,j) \in \mathcal{I}} \left\{ \left| (P_n - P)(y \varphi_{I_k,j}(x)) \right| \leq L_{A,A_+,r,c_{\min},c_{M,\mathrm{Leb}},\gamma} \sqrt{\frac{\ln n}{n}} \right\}$$

we deduce that

$$\mathbb{P}\left(\Omega_n^{(1)}\right) \geq 1 - 2Dn^{-\gamma} \ . \tag{3.87}$$

Hence the expected bound (3.84) holds on $\Omega_n^{(1)}$, for all $n \geq 1$.
We turn now to the proof of inequality (3.83). Let the index $(I_k, j) \in \mathcal{I}$ be fixed. By Cauchy-Schwarz inequality, we have

$$\sum_{i=0}^{r} |(P_n - P)(\varphi_{I_k,j}\varphi_{I_k,i})| \leq \sqrt{r+1}\sqrt{\sum_{i=0}^{r} ((P_n - P)(\varphi_{I_k,j}\varphi_{I_k,i}))^2} \ . \tag{3.88}$$

Let write

$$\chi_{I_k,j} = \sqrt{\sum_{i=0}^{r} ((P_n - P)(\varphi_{I_k,j}\varphi_{I_k,i}))^2} \ \text{ and } \ B_{I_k} = \left\{ \sum_{i=0}^{r} \beta_{I_k,i}\varphi_{I_k,i} \ ; \ (\beta_{I_k,i})_{i=0}^{r} \in \mathbb{R}^{r+1} \text{ and } \sum_{i=0}^{r} \beta_{I_k,i}^2 \leq 1 \right\} \ .$$

By Cauchy-Schwarz inequality again, it holds

$$\chi_{I_k,j} = \sup_{s \in B_{I_k}} |(P_n - P)(\varphi_{I_k,j}s)| \ .$$

Then, Bousquet's inequality (7.47), applied with $\varepsilon = 1$ and $\mathcal{F} = B_{I_k}$, implies that

$$\mathbb{P}\left[\chi_{I_k,j} - \mathbb{E}[\chi_{I_k,j}] \geq \sqrt{2\sigma_{I_k,j}^2 \frac{x}{n}} + \mathbb{E}[\chi_{I_k,j}] + \frac{4}{3}\frac{b_{I_k,j}x}{n}\right] \leq \exp(-x) \tag{3.89}$$

where, by (3.68),

$$\sigma_{I_k,j}^2 = \sup_{s \in B_{I_k}} \mathrm{Var}(\varphi_{I_k,j}s) \leq \|\varphi_{I_k,j}\|_\infty^2 \leq \frac{L_{r,c_{\min}}}{\mathrm{Leb}(I_k)} \tag{3.90}$$

and

$$b_{I_k,j} \leq 2 \sup_{s \in B_{I_k}} \|\varphi_{I_k,j}s\|_\infty \leq 2\|\varphi_{I_k,j}\|_\infty \sup_{s \in B_{I_k}} \|s\|_\infty \ . \tag{3.91}$$

Moreover, for $s = \sum_{i=0}^{r} \beta_{I_k,i}\varphi_{I_k,i} \in B_{I_k}$, we have $\max_i |\beta_{I_k,i}| \leq \sqrt{\sum_{i=0}^{r} \beta_{I_k,i}^2} \leq 1$, so by (3.68),

$$\sup_{s \in B_{I_k}} \|s\|_\infty \leq \sum_{i=0}^{r} \|\varphi_{I_k,i}\|_\infty \leq \frac{L_{r,c_{\min}}}{\sqrt{\mathrm{Leb}(I_k)}}$$

and injecting the last bound in (3.91) we get

$$b_{I_k,j} \leq \|\varphi_{I_k,j}\|_\infty \frac{L_{r,c_{\min}}}{\sqrt{\mathrm{Leb}(I_k)}} \leq \frac{L_{r,c_{\min}}}{\mathrm{Leb}(I_k)} \ . \tag{3.92}$$

In addition, we have

$$\begin{aligned}
\mathbb{E}[\chi_{I_k,j}] &\leq \sqrt{\mathbb{E}\left[\chi_{I_k,j}^2\right]} = \sqrt{\frac{\sum_{i=0}^{r} \mathrm{Var}(\varphi_{I_k,j}\varphi_{I_k,i})}{n}} \\
&\leq \|\varphi_{I_k,j}\|_\infty \sqrt{\frac{\sum_{i=0}^{r} P\left(\varphi_{I_k,i}^2\right)}{n}} \\
&= \|\varphi_{I_k,j}\|_\infty \sqrt{\frac{r+1}{n}} \\
&\leq L_{r,c_{\min}} \sqrt{\frac{1}{n\,\mathrm{Leb}(I_k)}} \ .
\end{aligned} \tag{3.93}$$

Therefore, combining (3.90), (3.92), (3.93) and (3.89) while taking $x = \gamma \ln n$, we get

$$\mathbb{P}\left[\chi_{I_k,j} \geq L_{r,c_{\min},\gamma}\left(\sqrt{\frac{1}{n\,\mathrm{Leb}\,(I_k)}} + \sqrt{\frac{\ln n}{n\,\mathrm{Leb}\,(I_k)}} + \frac{\ln n}{n\,\mathrm{Leb}\,(I_k)}\right)\right] \leq n^{-\gamma}\ . \tag{3.94}$$

Now, since by (3.49) and the fact that $D \leq A_+ n\,(\ln n)^{-2}$ we have

$$\frac{1}{\mathrm{Leb}\,(I_k)} \leq c_{M,\mathrm{Leb}}^{-2} D \leq c_{M,\mathrm{Leb}}^{-2} A_+ \frac{n}{(\ln n)^2}\ ,$$

we obtain from (3.94) that a positive constant $L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}$ exists, depending only on $\gamma, r, A_+, c_{\min}$ and $c_{M,\mathrm{Leb}}$ such that

$$\mathbb{P}\left[\chi_{I_k,j} \geq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n\,\mathrm{Leb}\,(I_k)}}\right] \leq n^{-\gamma}\ . \tag{3.95}$$

Finally, define

$$\Omega_n^{(2)} = \bigcap_{(I_k,j)\in\mathcal{I}} \left\{\chi_{I_k,j} \leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n\,\mathrm{Leb}\,(I_k)}}\right\}\ .$$

For all $n \geq n_0\,(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma)$, we have

$$\begin{aligned}
&\sqrt{r+1} \times L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{\ln n}{n\,\mathrm{Leb}\,(I_k)}} \\
&\leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{D\ln n}{n}} \\
&\leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\frac{1}{\sqrt{\ln n}} \leq \frac{1}{2}\ .
\end{aligned} \tag{3.96}$$

Moreover by (3.95) it holds

$$\mathbb{P}\left(\Omega_n^{(2)}\right) \geq 1 - Dn^{-\gamma} \tag{3.97}$$

and, by (3.88), the expected bound (3.83) holds on $\Omega_n^{(2)}$, for all $n \geq n_0\,(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma)$. Next, notice that for a $D \times D$ matrix $L = \left(L_{(I_k,j),(I_l,i)}\right)_{(I_k,j),(I_l,i)\in\mathcal{I}\times\mathcal{I}}$ we have the following classical formula,

$$\|L\| = \max_{(I_k,j)\in\mathcal{I}} \sum_{(I_l,i)\in\mathcal{I}} \left|L_{(I_k,j),(I_l,i)}\right|\ .$$

Applied to the matrix of interest $L_{n,D}$, this gives

$$\begin{aligned}
\|L_{n,D}\| &= \max_{(I_k,j)\in\mathcal{I}} \sum_{(I_l,i)\in\mathcal{I}} \left|(P_n - P)\left(\varphi_{I_k,j}\varphi_{I_l,i}\right)\right| \\
&= \max_{k\in\{0,\ldots,m-1\}} \max_{j\in\{0,\ldots,r\}} \left\{\sum_{(I_l,i)\in\mathcal{I}} \left|(P_n - P)\left(\varphi_{I_k,j}\varphi_{I_l,i}\right)\right|\right\}\ .
\end{aligned} \tag{3.98}$$

Thus, using formula (3.98), inequalities (3.83), (3.49) and (3.96) give that for all $n \geq n_0\,(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma)$, we have on $\Omega_n^{(2)}$,

$$\|L_{n,D}\| \leq L_{r,A_+,c_{\min},c_{M,\mathrm{Leb}},\gamma}\sqrt{\frac{D\ln n}{n}} \leq \frac{1}{2}\ .$$

Finally, by setting $\Omega_n = \Omega_n^{(1)} \bigcap \Omega_n^{(2)}$, we have $\mathbb{P}\left(\Omega_n\right) \geq 1 - 3Dn^{-\gamma}$, and inequalities (3.83), (3.82) and (3.84) are satisfied on Ω_n for all $n \geq n_0\,(r, A_+, c_{\min}, c_{M,\mathrm{Leb}}, \gamma)$, which completes the proof of Lemma 3.6. ■

3.6.3 Proofs of Section 3.3

In order to express the quantities of interest in the proofs of Theorems 3.1 and 3.2, we need preliminary definitions. As usual, M is a linear space of finite dimension D. Furthermore, let $\alpha > 0$ be fixed and for $R_{n,D,\alpha}$ defined in (**H5**), see Section 3.3.1, we set

$$\tilde{R}_{n,D,\alpha} = \max\left\{ R_{n,D,\alpha} \; ; \; A_\infty \sqrt{\frac{D \ln n}{n}} \right\} \tag{3.99}$$

where A_∞ is a positive constant to be chosen later. Moreover, we set

$$\nu_n = \max\left\{ \sqrt{\frac{\ln n}{D}} \; ; \; \sqrt{\frac{D \ln n}{n}} \; ; \; R_{n,D,\alpha} \right\} \; . \tag{3.100}$$

Following then heuristics given in Section 7.3.2 of Chapter 7, our analysis is localized in the subset

$$B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right) = \left\{ s \in M, \|s - s_M\|_\infty \leq \tilde{R}_{n,D,\alpha} \right\}$$

of M.
Let us define several slices of excess risk on the model M : for any $C \geq 0$,

$$\mathcal{F}_C = \{ s \in M, P\left(Ks - Ks_M\right) \leq C \} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right)$$
$$\mathcal{F}_{>C} = \{ s \in M, P\left(Ks - Ks_M\right) > C \} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right)$$

and for any interval $J \subset \mathbb{R}$,

$$\mathcal{F}_J = \{ s \in M, P\left(Ks - Ks_M\right) \in J \} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right) \; .$$

We also define, for all $L \geq 0$,

$$D_L = \{ s \in M, P\left(Ks - Ks_M\right) = L \} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right) \; .$$

Recall that, by Lemma 3.1 of Section 3.2.2, the contrasted functions satisfy, for every $s \in M$ and $z = (x, y) \in \mathcal{X} \times \mathbb{R}$,

$$\left(Ks\right)(z) - \left(Ks_M\right)(z) = \psi_{1,M}(z)\left(s - s_M\right)(x) + \psi_2\left(\left(s - s_M\right)(x)\right)$$

where $\psi_{1,M}(z) = -2\left(y - s_M(x)\right)$ and $\psi_2(t) = t^2$, for all $t \in \mathbb{R}$. For convenience, we will use the following notation, for any $s \in M$,

$$\psi_2 \circ \left(s - s_M\right) : x \in \mathcal{X} \longmapsto \psi_2\left(\left(s - s_M\right)(x)\right) \; .$$

Note that, for all $s \in M$,

$$P\left(\psi_{1,M} \cdot s\right) = 0 \tag{3.101}$$

and by (**H1**) inequality (3.15) holds true, that is

$$\|\psi_{1,M}\|_\infty \leq 4A \; . \tag{3.102}$$

Also, for $\mathcal{K}_{1,M}$ defined in Section 3.3.3, we have

$$\mathcal{K}_{1,M} = \sqrt{\frac{1}{D}\sum_{k=1}^{D} \mathrm{Var}\left(\psi_{1,M} \cdot \varphi_k\right)}$$

for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$. Moreover, inequality (3.31) holds under (**H1**) and we have

$$\mathcal{K}_{1,M} \leq 2\sigma_{\max} + 4A \leq 6A \ . \tag{3.103}$$

Assuming (**H2**), we have from (3.32)

$$0 < 2\sigma_{\min} \leq \mathcal{K}_{1,M} \ . \tag{3.104}$$

Finally, when (**H3**) holds (it is the case when (**H4**) holds), we have by (3.16),

$$\sup_{s \in M,\ \|s\|_2 \leq 1} \|s\|_\infty \leq A_{3,M}\sqrt{D} \tag{3.105}$$

and so, for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$, it holds for all $k \in \{1, ..., D\}$, as $P\left(\varphi_k^2\right) = 1$,

$$\|\varphi_k\|_\infty \leq A_{3,M}\sqrt{D} \ . \tag{3.106}$$

Proofs of the theorems

The proof of Theorem 3.1 relies on Lemmas 3.12, 3.13 and 3.14 stated in Section 3.6.4, and that give sharp estimates of suprema of the empirical process on the constrasted functions over slices of interest.

Proof of Theorem 3.1. Let $\alpha > 0$ be fixed and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying (**H4**). We divide the proof of Theorem 3.1 into four parts, corresponding to the four Inequalities (3.22), (3.23), (3.24) and (3.25). The values of A_0 and A_∞, respectively defined in (3.21) and (3.99), will then be chosen at the end of the proof.

Proof of Inequality (3.22). Let $r \in (1, 2]$ to be chosen later and $C > 0$ such that

$$rC = \frac{D}{4n}\mathcal{K}_{1,M}^2 \ . \tag{3.107}$$

By (**H5**) there exists a positive integer n_1 such that it holds, for all $n \geq n_1$,

$$\mathbb{P}\left(P\left(Ks_n - Ks_M\right) \leq C\right) \leq \mathbb{P}\left(\left\{P\left(Ks_n - Ks_M\right) \leq C\right\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} \tag{3.108}$$

and also

$$\begin{aligned}
&\mathbb{P}\left(\left\{P\left(Ks_n - Ks_M\right) \leq C\right\} \bigcap \Omega_{\infty,\alpha}\right) \\
&\leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C} P_n\left(Ks - Ks_M\right) \leq \inf_{s \in \mathcal{F}_{>C}} P_n\left(Ks - Ks_M\right)\right) \\
&\leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C} P_n\left(Ks - Ks_M\right) \leq \inf_{s \in \mathcal{F}_{(C,rC]}} P_n\left(Ks - Ks_M\right)\right) \\
&= \mathbb{P}\left(\sup_{s \in \mathcal{F}_C} P_n\left(Ks_M - Ks\right) \geq \sup_{s \in \mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right)\right).
\end{aligned} \tag{3.109}$$

Now, by (3.107) and (3.104) we have

$$\frac{D}{2n}\sigma_{\min}^2 \leq C \leq \left(1 + A_4\nu_n\right)^2 \frac{D}{4n}\mathcal{K}_{1,M}^2$$

where A_4 is defined in Lemma 3.12. Hence we can apply Lemma 3.12 with $\alpha = \beta$, $A_l = \sigma_{\min}^2/2$ and $A_{3,M} = r_M(\varphi)$, by Remark 3.1. Therefore it holds, for all $n \geq n_0(A_\infty, A_{cons}, A_+, \sigma_{\min}, \alpha)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C} P_n(Ks_M - Ks) \geq \left(1 + L_{A_\infty,A,r_M(\varphi),\sigma_{\min},A_-,\alpha} \times \nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq 2n^{-\alpha} . \tag{3.110}$$

Moreover, by using (3.104) and (3.103) in (3.107) we get

$$\frac{D}{n}\sigma_{\min}^2 \leq rC \leq \frac{D}{n}(\sigma_{\max} + 2A)^2 .$$

We then apply Lemma 3.14 with

$$\alpha = \beta,\ A_l = \sigma_{\min}^2,\ A_u = (\sigma_{\max} + 2A)^2$$

and

$$A_\infty \geq 64\sqrt{2}B_2 A(\sigma_{\max} + 2A)\sigma_{\min}^{-1} r_M(\varphi) , \tag{3.111}$$

so it holds for all $n \geq n_0(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, \alpha)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n(Ks_M - Ks) \leq \left(1 - L_{A_-,A,A_\infty,\sigma_{\max},\sigma_{\min},r_M(\varphi),\alpha} \times \nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq 2n^{-\alpha} . \tag{3.112}$$

Now, from (3.110) and (3.112) we can find a positive constant $\tilde{A}_0$, only depending on A_-, A, A_∞, $\sigma_{\max}, \sigma_{\min}, r_M(\varphi)$ and α, such that for all $n \geq n_0(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, \alpha)$, there exists an event of probability at least $1 - 4n^{-\alpha}$ on which

$$\sup_{s\in\mathcal{F}_C} P_n(Ks_M - Ks) \leq \left(1 + \tilde{A}_0\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C \tag{3.113}$$

and

$$\sup_{s\in\mathcal{F}_{(C,rC]}} P_n(Ks_M - Ks) \geq \left(1 - \tilde{A}_0\nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC . \tag{3.114}$$

Hence, from (3.113) and (3.114) we deduce, using (3.108) and (3.109), that if we choose $r \in (1, 2]$ such that

$$\left(1 + \tilde{A}_0\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C < \left(1 - \tilde{A}_0\nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC \tag{3.115}$$

then, for all $n \geq n_0(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M(\varphi), \sigma_{\max}, \sigma_{\min}, n_1, \alpha)$ we have

$$P(Ks_n - Ks_M) \geq C$$

with probability at least $1 - 5n^{-\alpha}$. Now, by (3.107) it holds

$$\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} = 2rC = \frac{1}{2}\frac{D}{n}\mathcal{K}_{1,M}^2 ,$$

and as a consequence Inequality (3.115) is equivalent to

$$\left(1 - 2\tilde{A}_0\nu_n\right)r - 2\left(1 + \tilde{A}_0\nu_n\right)\sqrt{r} + 1 > 0 . \tag{3.116}$$

Moreover, we have by (3.100) and (**H5**), for all $n \geq n_0\left(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha\right)$,

$$\tilde{A}_0\nu_n \leq \frac{1}{4} \tag{3.117}$$

and so, for all $n \geq n_0\left(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha\right)$, simple computations involving (3.117) show that by taking

$$r = 1 + 48\sqrt{\tilde{A}_0 \nu_n} \tag{3.118}$$

inequality (3.116) is satisfied. Notice that, for all $n \geq n_0\left(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha\right)$ we have $0 < 48\sqrt{\tilde{A}_0 \nu_n} < 1$, so that $r \in (1,2)$. Finally, we compute C by (3.107) and (3.118), in such a way that for all $n \geq n_0\left(A_+, A_-, A_{cons}, \tilde{A}_0, \alpha\right)$,

$$C = \frac{rC}{r} = \frac{1}{1 + 48\sqrt{\tilde{A}_0 \nu_n}} \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq \left(1 - 48\sqrt{\tilde{A}_0 \nu_n}\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 > 0 \tag{3.119}$$

which yields the result by noticing that the dependence on $\sigma_{\max}$ can be released in n_0 and $\tilde{A}_0$ since by (**H1**) we have $\sigma_{\max} \leq A$.

Proof of Inequality (3.23). Let $C > 0$ and $\delta \in \left(0, \frac{1}{2}\right)$ to be chosen later in such a way that

$$(1 - \delta)\, C = \frac{D}{4n} \mathcal{K}_{1,M}^2 \tag{3.120}$$

and

$$C \geq \frac{1}{4} \left(1 + A_5 \nu_n\right)^2 \frac{D}{n} \mathcal{K}_{1,M}^2 \ , \tag{3.121}$$

where A_5 is defined in Lemma 3.13. We have by (**H5**), for all $n \geq n_1$,

$$\mathbb{P}\left(P\left(Ks_n - Ks_M\right) > C\right) \leq \mathbb{P}\left(\left\{P\left(Ks_n - Ks_M\right) > C\right\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} \tag{3.122}$$

and also

$$\begin{aligned}
&\mathbb{P}\left(\left\{P\left(Ks_n - Ks_M\right) > C\right\} \bigcap \Omega_{\infty,\alpha}\right) \\
&\leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C} P_n\left(Ks - Ks_M\right) \geq \inf_{s \in \mathcal{F}_{>C}} P_n\left(Ks - Ks_M\right)\right) \\
&= \mathbb{P}\left(\sup_{s \in \mathcal{F}_C} P_n\left(Ks_M - Ks\right) \leq \sup_{s \in \mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right)\right) \\
&\leq \mathbb{P}\left(\sup_{s \in \mathcal{F}_{\left(\frac{C}{2},(1-\delta)C\right]}} P_n\left(Ks_M - Ks\right) \leq \sup_{s \in \mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right)\right) \ .
\end{aligned} \tag{3.123}$$

Now by (3.121) we can apply Lemma 3.13 with $\alpha = \beta$ and we obtain, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, \alpha\right)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \left(1 + A_5 \nu_n\right) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C\right] \leq 2n^{-\alpha} \tag{3.124}$$

where A_5 only depends on $A, A_{3,M}, A_\infty, \sigma_{\min}, A_-$ and α. Moreover, we can take $A_{3,M} = r_M(\varphi)$ by Remark 3.1. Also, by (3.120), (3.104) and (3.103) we can apply Lemma 3.14 with the quantity C in Lemma 3.14 replaced by $C/2$, $\alpha = \beta$, $r = 2(1-\delta)$, $A_u = \left(\sigma_{\max} + 2A\right)^2$, $A_l = \sigma_{\min}^2$ and the constant A_∞ satisfying

$$A_\infty \geq 64\sqrt{2} B_2 A \left(\sigma_{\max} + 2A\right) \sigma_{\min}^{-1} r_M(\varphi) \ , \tag{3.125}$$

and so it holds, for all $n \geq n_0\left(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M\left(\varphi\right), \sigma_{\max}, \sigma_{\min}, \alpha\right)$,

$$\mathbb{P}\left(\begin{array}{c}\sup_{s\in\mathcal{F}_{\left[\frac{C}{2},(1-\delta)C\right]}} P_n\left(Ks_M - Ks\right) \\ \leq \left(1 - L_{A_-,A,A_\infty,\sigma_{\max},\sigma_{\min},r_M(\varphi),\alpha} \times \nu_n\right)\sqrt{\frac{(1-\delta)CD}{n}}\mathcal{K}_{1,M} - (1-\delta)C\end{array}\right) \leq 2n^{-\alpha}\ . \quad (3.126)$$

Hence from (3.124) and (3.126), we deduce that a positive constant $\check{A}_0$ exists, only depending on $A_-, A, A_\infty, \sigma_{\max}, \sigma_{\min}, r_M\left(\varphi\right)$ and α, such that
for all $n \geq n_0\left(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M\left(\varphi\right), \sigma_{\max}, \sigma_{\min}, \alpha\right)$ it holds on an event of probability at least $1 - 4n^{-\alpha}$,

$$\sup_{s\in\mathcal{F}_{\left[\frac{C}{2},(1-\delta)C\right]}} P_n\left(Ks_M - Ks\right) \geq \left(1 - \check{A}_0\nu_n\right)\sqrt{\frac{(1-\delta)CD}{n}}\mathcal{K}_{1,M} - (1-\delta)C \quad (3.127)$$

and

$$\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \leq \left(1 + \check{A}_0\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\ . \quad (3.128)$$

Now, from (3.127) and (3.128) we deduce, using (3.122) and (3.123), that if we choose $\delta \in \left(0, \frac{1}{2}\right)$ such that (3.121) and

$$\left(1 + \check{A}_0\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C < \left(1 - \check{A}_0\nu_n\right)\sqrt{\frac{(1-\delta)CD}{n}}\mathcal{K}_{1,M} - (1-\delta)C \quad (3.129)$$

are satisfied then, for all $n \geq n_0\left(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M\left(\varphi\right), \sigma_{\max}, \sigma_{\min}, n_1, \alpha\right)$,

$$P\left(Ks_n - Ks_M\right) \leq C\ ,$$

with probability at least $1 - 5n^{-\alpha}$. By (3.120) it holds

$$\sqrt{\frac{(1-\delta)CD}{n}}\mathcal{K}_{1,M} = 2\left(1-\delta\right)C = \frac{1}{2}\frac{D}{n}\mathcal{K}_{1,M}^2\ ,$$

and by consequence, inequality (3.129) is equivalent to

$$\left(1 - 2\check{A}_0\nu_n\right)\left(1-\delta\right) - 2\left(1 + \check{A}_0\nu_n\right)\sqrt{1-\delta} + 1 > 0\ . \quad (3.130)$$

Moreover, we have by (3.100) and (**H5**), for all $n \geq n_0\left(A_+, A_-, A_{cons}, \check{A}_0, A_5, \alpha\right)$,

$$\left(\check{A}_0 \vee A_5\right)\nu_n < \frac{1}{72} \quad (3.131)$$

and so, for all $n \geq n_0\left(A_+, A_-, A_{cons}, \check{A}_0, \alpha\right)$, simple computations involving (3.131) show that by taking

$$\delta = 6\left(\sqrt{\check{A}_0} \vee \sqrt{A_5}\right)\sqrt{\nu_n}\ , \quad (3.132)$$

inequalities (3.130) and (3.121) are satisfied and $\delta \in \left(0, \frac{1}{2}\right)$. Finally, we can compute C by (3.120) and (3.132), in such a way that for all $n \geq n_0\left(A_+, A_-, A_{cons}, \check{A}_0, \alpha\right)$

$$0 < C = \frac{(1-\delta)C}{(1-\delta)} = \frac{1}{(1-\delta)}\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \leq \left(1 + 12\left(\sqrt{\check{A}_0} \vee \sqrt{A_5}\right)\sqrt{\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\ , \quad (3.133)$$

which yields the result by noticing that the dependence on $\sigma_{\max}$ can be released from n_0 and $\check{A}_0$ since by (**H1**) we have $\sigma_{\max} \leq A$.

Proof of Inequality (3.24). Let $C = \frac{D}{8n}\mathcal{K}_{1,M}^2 > 0$ and let $r = 2$. By (3.103) and (3.104) we have

$$\frac{D}{n}\sigma_{\min}^2 \leq rC = \frac{D}{4n}\mathcal{K}_{1,M}^2 \leq \frac{D}{n}\left(\sigma_{\max} + 2A\right)^2$$

so we can apply Lemma 3.14 with $\alpha = \beta$, $A_l = \sigma_{\min}^2$ and $A_u = (\sigma_{\max} + 2A)^2$. So if

$$A_\infty \geq 64\sqrt{2}B_2 A\left(\sigma_{\max} + 2A\right)\sigma_{\min}^{-1} r_M\left(\varphi\right) \ , \tag{3.134}$$

it holds, for all $n \geq n_0\left(A_-, A_+, A, A_\infty, A_{cons}, B_2, r_M\left(\varphi\right), \sigma_{\max}, \sigma_{\min}, \alpha\right)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_-,A,A_\infty,\sigma_{\max},\sigma_{\min},r_M(\varphi),\alpha} \times \nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq 2n^{-\alpha} \ . \tag{3.135}$$

Since $rC = \frac{D}{4n}\mathcal{K}_{1,M}^2$, if we set $\hat{A}_0 = 2L_{A_-,A,A_\infty,\sigma_{\max},\sigma_{\min},r_M(\varphi),\alpha}$ with $L_{A_-,A,A_\infty,\sigma_{\max},\sigma_{\min},r_M(\varphi),\alpha}$ the constant in (3.135), we get

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - \hat{A}_0\nu_n\right)\frac{D}{4n}\mathcal{K}_{1,M}^2\right) \leq 2n^{-\alpha} \ . \tag{3.136}$$

Notice that

$$P_n\left(Ks_M - Ks_n\right) = \sup_{s\in M} P_n\left(Ks_M - Ks\right) \geq \sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right)$$

so from (3.136) we deduce that

$$\mathbb{P}\left(P_n\left(Ks_M - Ks_n\right) \geq \left(1 - \hat{A}_0\nu_n\right)\frac{D}{4n}\mathcal{K}_{1,M}^2\right) \geq 1 - 2n^{-\alpha} \ . \tag{3.137}$$

Remark 3.2 *Notice that in the proof of inequality (**3.24**), we do not need to assume the consistency of the least-squares estimator s_n towards the projection s_M. Straightforward adaptations of Lemma 3.14 allow to take*

$$\tilde{\nu}_n = \max\left\{\sqrt{\frac{\ln n}{D}},\ \sqrt{\frac{D\ln n}{n}}\right\}$$

instead of the quantity ν_n defined in (3.100). This readily gives the expected bound (3.26) of Theorem 3.1.

Proof of Inequality (3.25). Let

$$C = \frac{1}{4}\left(1 + A_5\nu_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2 > 0 \tag{3.138}$$

where A_5 is defined in Lemma 3.13 applied with $\beta = \alpha$. By (**H5**) we have

$$\mathbb{P}\left(P_n\left(Ks_M - Ks_n\right) > C\right) \leq \mathbb{P}\left(\left\{P_n\left(Ks_M - Ks_n\right) > C\right\}\bigcap\Omega_{\infty,\alpha}\right) + n^{-\alpha} \ . \tag{3.139}$$

Moreover, on $\Omega_{\infty,\alpha}$, we have

$$P_n\left(Ks_M - Ks_n\right) = \sup_{s \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right)} P_n\left(Ks_M - Ks\right) = \sup_{s \in \mathcal{F}_{>0}} P_n\left(Ks_M - Ks\right) \tag{3.140}$$

and by (3.198) of Lemma 3.13 applied with $\alpha = \beta$ it holds, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, \alpha\right)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{>0}} P_n\left(Ks_M - Ks\right) > C\right) \leq 2n^{-\alpha} \ . \tag{3.141}$$

Finally, using (3.140) and (3.141) in (3.139) we get, for all $n \geq n_0\left(A_\infty, A_{cons}, n_1, A_+, \alpha\right)$,

$$\mathbb{P}\left(P_n\left(Ks_M - Ks_n\right) > C\right) \leq 3n^{-\alpha} \ .$$

Conclusion. To complete the proof of Theorem 3.1, just notice that by (3.111), (3.125) and (3.134) we can take

$$A_\infty = 64\sqrt{2}B_2 A\left(\sigma_{\max} + 2A\right)\sigma_{\min}^{-1} r_M\left(\varphi\right)$$

and by (3.119), (3.133), (3.137) and (3.138),

$$A_0 = \max\left\{48\sqrt{\tilde{A}_0},\ 12\left(\sqrt{\check{A}_0} \vee \sqrt{A_5}\right),\ \sqrt{\hat{A}_0},\ \sqrt{A_5}\right\}$$

is convenient. ■

Proof of Theorem 3.2. We localize our analysis in the subset

$$B_{(M,L_\infty)}\left(s_M, R_{n,D,\alpha}\right) = \left\{s \in M, \left\|s - s_M\right\|_\infty \leq R_{n,D,\alpha}\right\} \subset M \ .$$

Unlike in the proof of Theorem 3.1, see (3.99), we need not to consider the quantity $\tilde{R}_{n,D,\alpha}$, a radius possibly larger than $R_{n,D,\alpha}$. Indeed, the use of $\tilde{R}_{n,D,\alpha}$ rather than $R_{n,D,\alpha}$ in the proof of Theorem 3.1 is only needed in Lemma 3.8, where we derive a sharp lower bound for the mean of the supremum of the empirical process indexed by the contrasted functions centered by the contrasted projection over a slice of interest. To prove Theorem 3.2, we just need upper bounds, and Lemma 3.8 is avoided as well as the use of $\tilde{R}_{n,D,\alpha}$.
Let us define several slices of excess risk on the model M : for any $C \geq 0$,

$$\mathcal{G}_C = \left\{s \in M, P\left(Ks - Ks_M\right) \leq C\right\} \bigcap B_{(M,L_\infty)}\left(s_M, R_{n,D,\alpha}\right) \ ,$$
$$\mathcal{G}_{>C} = \left\{s \in M, P\left(Ks - Ks_M\right) > C\right\} \bigcap B_{(M,L_\infty)}\left(s_M, R_{n,D,\alpha}\right) \ .$$

We also define, for all $U \geq 0$,

$$\mathcal{D}_U = \left\{s \in M, P\left(Ks - Ks_M\right) = U\right\} \bigcap B_{(M,L_\infty)}\left(s_M, R_{n,D,\alpha}\right) \ .$$

I. Proof of Inequality (3.27). Let $C_1 > 0$ to fixed later, satisfying

$$C_1 \geq \frac{D}{n} =: C_- > 0 \ . \tag{3.142}$$

We have by (**H5**), for all $n \geq n_1$,

$$\mathbb{P}\left(P\left(Ks_n - Ks_M\right) > C_1\right) \leq \mathbb{P}\left(\left\{P\left(Ks_n - Ks_M\right) > C_1\right\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} \tag{3.143}$$

and also

$$\begin{aligned}
&\mathbb{P}\left(\left\{P\left(Ks_n - Ks_M\right) > C_1\right\} \bigcap \Omega_{\infty,\alpha}\right) \\
&\leq \mathbb{P}\left(\inf_{s \in \mathcal{G}_{C_1}} P_n\left(Ks - Ks_M\right) \geq \inf_{s \in \mathcal{G}_{>C_1}} P_n\left(Ks - Ks_M\right)\right) \\
&= \mathbb{P}\left(\sup_{s \in \mathcal{G}_{C_1}} P_n\left(Ks_M - Ks\right) \leq \sup_{s \in \mathcal{G}_{>C_1}} P_n\left(Ks_M - Ks\right)\right) \\
&\leq \mathbb{P}\left(0 \leq \sup_{s \in \mathcal{G}_{>C_1}} P_n\left(Ks_M - Ks\right)\right) .
\end{aligned} \tag{3.144}$$

Moreover, it holds

$$\begin{aligned}
&\sup_{s \in \mathcal{G}_{>C_1}} P_n\left(Ks_M - Ks\right) \\
&= \sup_{s \in \mathcal{G}_{>C_1}} \left\{P_n\left(\psi_{1,M} \cdot \left(s_M - s\right) - \psi_2 \circ \left(s - s_M\right)\right)\right\} \\
&= \sup_{s \in \mathcal{G}_{>C_1}} \left\{\left(P_n - P\right)\left(\psi_{1,M} \cdot \left(s_M - s\right)\right) - \left(P_n - P\right)\left(\psi_2 \circ \left(s - s_M\right)\right) - P\left(Ks - Ks_M\right)\right\} \\
&= \sup_{s \in \mathcal{G}_{>C_1}} \left\{\left(P_n - P\right)\left(\psi_{1,M} \cdot \left(s_M - s\right)\right) - P\left(Ks - Ks_M\right) - \left(P_n - P\right)\left(\psi_2 \circ \left(s - s_M\right)\right)\right\} \\
&= \sup_{U > C_1} \sup_{s \in \mathcal{D}_U} \left\{\left(P_n - P\right)\left(\psi_{1,M} \cdot \left(s_M - s\right)\right) - U - \left(P_n - P\right)\left(\psi_2 \circ \left(s - s_M\right)\right)\right\} \\
&\leq \sup_{U > C_1} \left\{\sqrt{U}\sqrt{\sum_{k=1}^{D} \left(P_n - P\right)^2\left(\psi_{1,M} \cdot \varphi_k\right)} - U + \sup_{s \in \mathcal{G}_U} \left|\left(P_n - P\right)\left(\psi_2 \circ \left(s - s_M\right)\right)\right|\right\} .
\end{aligned} \tag{3.145}$$

Now, from inequality (3.164) of Lemma 3.7 applied with $\beta = \alpha$, we get

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D} \left(P_n - P\right)^2\left(\psi_{1,M} \cdot \varphi_k\right)} \geq L_{A,A_{3,M},\alpha}\sqrt{\frac{D \vee \ln n}{n}}\right] \leq n^{-\alpha} . \tag{3.146}$$

In addition, we handle the empirical process indexed by the second order terms by straightforward modifications of Lemmas 3.10 and 3.11 as well as their proofs. It thus holds, by the same type of arguments as those given in Lemma 3.10,

$$\mathbb{E}\left[\sup_{s \in \mathcal{G}_{C_1}} \left|\left(P_n - P\right)\left(\psi_{2,M}^s \cdot \left(s - s_M\right)\right)\right|\right] \leq 8\sqrt{\frac{CD}{n}} R_{n,D,\alpha} . \tag{3.147}$$

Moreover, using (3.147), the same type of arguments as those leading to inequality (3.191) of Lemma 3.11, allow to show that for any $q \geq 1$ and $j \in \mathbb{N}^*$, for all $x > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{s \in \mathcal{G}_{q^j C_-}} \left|\left(P_n - P\right)\left(\psi_2 \circ \left(s - s_M\right)\right)\right| \geq 16\sqrt{\frac{q^j C_- D}{n}} R_{n,D,\alpha} + \sqrt{\frac{2R_{n,D,\alpha}^2 q^j C_- x}{n}} + \frac{8}{3}\frac{R_{n,D,\alpha}^2 x}{n}\right] \\
&\leq \exp\left(-x\right) .
\end{aligned} \tag{3.148}$$

Hence, taking $x = \gamma \ln n$ in (3.148) and using the fact that $C_- = Dn^{-1} \geq n^{-1}$, we get

$$\mathbb{P}\left[\sup_{s\in\mathcal{G}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_{cons},\gamma} R_{n,D,\alpha} \sqrt{\frac{q^j C_- (D \vee \ln n)}{n}}\right] \leq n^{-\gamma} . \quad (3.149)$$

Now, by straightforward modifications of the proof of Lemma 3.11, we get that for all $n \geq n_0(A_{cons})$,

$$\mathbb{P}\left[\forall U > C_-, \quad \sup_{s\in\mathcal{G}_U} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_{cons},\alpha} R_{n,D,\alpha} \sqrt{\frac{U (D \vee \ln n)}{n}}\right] \geq 1 - n^{-\alpha} . \quad (3.150)$$

Combining (3.145), (3.146) and (3.150), we have on an event of probability at least $1 - 2n^{-\alpha}$, for all $n \geq n_0(A_{cons})$,

$$\begin{aligned}\sup_{s\in\mathcal{G}_{>C_1}} P_n(Ks_M - Ks) &\leq \sup_{U>C_1}\left\{L_{A,A_{3,M},\alpha}\sqrt{\frac{U(D\vee\ln n)}{n}} - U + L_{A_{cons},\alpha}R_{n,D,\alpha}\sqrt{\frac{U(D\vee\ln n)}{n}}\right\} \\ &\leq \sup_{U>C_1}\left\{L_{A,A_{cons},A_{3,M},\alpha}(1 + R_{n,D,\alpha})\sqrt{\frac{U(D\vee\ln n)}{n}} - U\right\} . \quad (3.151)\end{aligned}$$

Now, as $R_{n,D,\alpha} \leq A_{cons}(\ln n)^{-1/2}$, we deduce from (3.151) that for

$$C_1 = L_{A,A_{cons},A_{3,M},\alpha}\frac{D \vee \ln(n)}{n} > C_- \quad (3.152)$$

with $L_{A,A_{cons},A_{3,M},\alpha}$ large enough, it holds with probability at least $1 - 2n^{-\alpha}$ and for all $n \geq n_0(A_{cons})$,

$$\sup_{s\in\mathcal{G}_{>C_1}} P_n(Ks_M - Ks) < 0 ,$$

and so by using (3.143) and (3.144), this yields inequality (3.27).

II. Proof of Inequality (3.28). Let $C_2 > 0$ to fixed later, satisfying

$$C_2 \geq \frac{D}{n} = C_- > 0 . \quad (3.153)$$

We have by (**H5**), for all $n \geq n_1$,

$$\mathbb{P}(P_n(Ks_M - Ks_n) > C_2) \leq \mathbb{P}\left(\{P_n(Ks_M - Ks_n) > C_2\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} . \quad (3.154)$$

Moreover, we have on $\Omega_{\infty,\alpha}$,

$$\begin{aligned}P_n(Ks_M - Ks_n) &= \sup_{s\in B_{(M,L_\infty)}(s_M,R_{n,D,\alpha})} P_n(Ks_M - Ks) \\ &= \max\left\{\sup_{s\in\mathcal{G}_{C_1}} P_n(Ks_M - Ks) \; ; \; \sup_{s\in\mathcal{G}_{>C_1}} P_n(Ks_M - Ks)\right\} , \quad (3.155)\end{aligned}$$

where C_1 is defined in the first part of the proof dedicated to the establishment of inequality (3.27). Moreover, let us recall that in the first part of the proof, we have proved that an event of probability at least $1 - 2n^{-\alpha}$ exists, that we call Ω_1, such that it holds on this event, for all $n \geq n_0(A_{cons})$,

$$\sqrt{\sum_{k=1}^{D}(P_n - P)^2(\psi_{1,M}\cdot\varphi_k)} \leq L_{A,A_{3,M},\alpha}\sqrt{\frac{D\vee\ln n}{n}} , \quad (3.156)$$

$$\forall U > C_-, \quad \sup_{s \in \mathcal{G}_U} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_{cons},\alpha} R_{n,D,\alpha} \sqrt{\frac{U(D \vee \ln n)}{n}} , \tag{3.157}$$

and

$$\sup_{s \in \mathcal{G}_{>C_1}} P_n(Ks_M - Ks) < 0 . \tag{3.158}$$

By (3.155) and (3.158), we thus have on $\Omega_{\infty,\alpha} \bigcap \Omega_1$, for all $n \geq n_0(A_{cons})$,

$$0 \leq P_n(Ks_M - Ks_n) = \sup_{s \in \mathcal{G}_{C_1}} P_n(Ks_M - Ks) . \tag{3.159}$$

In addition, it holds

$$\begin{aligned}
&\sup_{s \in \mathcal{G}_{C_1}} P_n(Ks_M - Ks) \\
&= \sup_{s \in \mathcal{G}_{C_1}} \{P_n(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ (s - s_M))\} \\
&= \sup_{s \in \mathcal{G}_{C_1}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)(\psi_2 \circ (s - s_M)) - P(Ks - Ks_M)\} \\
&\leq \sup_{s \in \mathcal{G}_{C_1}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s))\} + \sup_{s \in \mathcal{G}_{C_1}} |(P_n - P)(\psi_2 \circ (s - s_M))| .
\end{aligned} \tag{3.160}$$

Now, we have on Ω_1, for all $n \geq n_0(A_{cons})$,

$$\begin{aligned}
\sup_{s \in \mathcal{G}_{C_1}} \{(P_n - P)(\psi_{1,M} \cdot (s_M - s))\} &\leq \sqrt{C_1} \sqrt{\sum_{k=1}^{D} (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \\
&\leq L_{A,A_{3,M},\alpha} \sqrt{\frac{C_1(D \vee \ln n)}{n}} \quad \text{by (3.156)} \\
&= L_{A,A_{cons},A_{3,M},\alpha} \frac{D \vee \ln(n)}{n} \quad \text{by (3.152)} ,
\end{aligned} \tag{3.161}$$

and also, by (3.157) and (3.152),

$$\begin{aligned}
\sup_{s \in \mathcal{G}_{C_1}} |(P_n - P)(\psi_2 \circ (s - s_M))| &\leq L_{A_{cons},\alpha} R_{n,D,\alpha} \sqrt{\frac{C_1(D \vee \ln n)}{n}} \\
&\leq L_{A,A_{cons},A_{3,M},\alpha} R_{n,D,\alpha} \frac{D \vee \ln(n)}{n} .
\end{aligned} \tag{3.162}$$

Finally, as $R_{n,D,\alpha} \leq A_{cons}(\ln n)^{-1/2}$, we deduce from (3.159), (3.160), (3.161) and (3.162), that it holds on $\Omega_{\infty,\alpha} \bigcap \Omega_1$, for all $n \geq n_0(A_{cons})$,

$$P_n(Ks_M - Ks_n) \leq L_{A,A_{cons},A_{3,M},\alpha} \frac{D \vee \ln(n)}{n} ,$$

and so, this yields to inequality (3.28) by using (3.154) and this concludes the proof of Theorem **3.2**. ■

3.6.4 Technical Lemmas

We state here some lemmas needed in the proofs of Section 3.6.3. First, in Lemmas 3.7, 3.8 and 3.9, we derive some controls, from above and from below, of the empirical process indexed by the "linear parts" of the contrasted functions over slices of interest. Secondly, we give upper bounds in Lemmas 3.10 and 3.11 for the empirical process indexed by the "quadratic parts" of the contrasted functions over slices of interest. And finally, we use all these results in Lemmas 3.12, 3.13 and 3.14 to derive upper and lower bounds for the empirical process indexed by the contrasted functions over slices of interest.

Lemma 3.7 *Assume that (**H1**), (**H2**) and (**H3**) hold. Then for any $\beta > 0$, by setting*

$$\tau_n = L_{A,A_{3,M},\sigma_{\min},\beta}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right) ,$$

It holds, for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|_2)$,

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D}\left(P_n - P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)} \geq (1+\tau_n)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} . \tag{3.163}$$

*If (**H1**) and (**H3**) hold, then for any $\beta > 0$, it holds*

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D}\left(P_n - P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)} \geq L_{A,A_{3,M},\beta}\sqrt{\frac{D\vee\ln n}{n}}\right] \leq n^{-\beta} . \tag{3.164}$$

Proof. By Cauchy-Schwarz inequality we have

$$\chi_M := \sqrt{\sum_{k=1}^{D}\left(P_n - P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)} = \sup_{s\in M,\ \|s\|_2\leq 1}\left\{\left|\left(P_n - P\right)\left(\psi_{1,M}\cdot s\right)\right|\right\} .$$

Hence, we get by Bousquet's inequality (7.48) applied with $\mathcal{F} = \{\psi_{1,M}\cdot s\ ;\ s\in M,\ \ \|s\|_2 \leq 1\}$, for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi_M \geq \sqrt{2\sigma^2\frac{x}{n}} + (1+\delta)\,\mathbb{E}\left[\chi_M\right] + \left(\frac{1}{3}+\frac{1}{\delta}\right)\frac{bx}{n}\right] \leq \exp\left(-x\right) \tag{3.165}$$

where

$$\sigma^2 \leq \sup_{s\in M,\ \|s\|_2\leq 1} P\left[\left(\psi_{1,M}\cdot s\right)^2\right] \leq \|\psi_{1,M}\|_\infty^2 \leq 16A^2 \qquad \text{by (3.102)}$$

and

$$b \leq \sup_{s\in M,\ \|s\|_2\leq 1}\left\|\psi_{1,M}\cdot s - P\left(\psi_{1,M}\cdot s\right)\right\|_\infty \leq 4A\sqrt{D}A_{3,M} \qquad \text{by (3.101), (3.102) and (3.105).}$$

Moreover,

$$\mathbb{E}\left[\chi_M\right] \leq \sqrt{\mathbb{E}\left[\chi_M^2\right]} = \sqrt{\frac{D}{n}}\mathcal{K}_{1,M} .$$

So, from (3.165) it follows that, for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi_M \geq \sqrt{32A^2\frac{x}{n}} + (1+\delta)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M} + \left(\frac{1}{3}+\frac{1}{\delta}\right)\frac{4A\sqrt{D}A_{3,M}x}{n}\right] \leq \exp\left(-x\right) . \tag{3.166}$$

Hence, taking $x = \beta\ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ in (3.166), we derive by (3.104) that a positive constant $L_{A,A_{3,M},\sigma_{\min},\beta}$ exists such that

$$\mathbb{P}\left[\chi_M \geq \left(1 + L_{A,A_{3,M},\sigma_{\min},\beta}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} ,$$

which yields inequality (3.163). By (3.103) we have $\mathcal{K}_{1,M} \leq 6A$, and by taking again $x = \beta \ln n$ and $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ in (3.166), simple computations give

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D} (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \geq L_{A,A_{3,M},\beta} \left(\sqrt{\frac{D}{n}} \vee \sqrt{\frac{\ln n}{n}} \vee \sqrt{\frac{D \ln n}{n^{3/2}}} \right)\right] \leq n^{-\beta} ,$$

and by consequence, (3.164) follows. ■

In the next lemma, we state sharp lower bounds for the mean of the supremum of the empirical process on the linear parts of constrasted functions of M belonging to a slice of excess risk. This is done for a model of reasonable dimension.

Lemma 3.8 *Let $r > 1$ and $C > 0$. Assume that (**H1**), (**H2**), (**H4**) and (3.17) hold and let $\varphi = (\varphi_k)_{k=1}^{D}$ be an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying (**H4**). If positive constants A_-, A_+, A_l, A_u exist such that*

$$A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2 \quad \text{and} \quad A_l \frac{D}{n} \leq rC \leq A_u \frac{D}{n} ,$$

and if the constant A_∞ defined in (3.99) satisfies

$$A_\infty \geq 64 B_2 A \sqrt{2A_u} \sigma_{\min}^{-1} r_M(\varphi) , \tag{3.167}$$

then a positive constant $L_{A,A_l,A_u,\sigma_{\min}}$ exists such that, for all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s))\right] \geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}\right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} . \tag{3.168}$$

Our argument leading to Lemma 3.8 shows that we have to assume that the constant A_∞ introduced in (3.99) is large enough. In order to prove Lemma 3.8 the following result is needed.

Lemma 3.9 *Let $r > 1$, $\beta > 0$ and $C \geq 0$. Assume that (**H1**), (**H2**), (**H4**) and (3.17) hold and let $\varphi = (\varphi_k)_{k=1}^{D}$ be an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying (**H4**). If positive constants A_+, A_- and A_u exist such that*

$$A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2 , \quad rC \leq A_u \frac{D}{n} ,$$

and if

$$A_\infty \geq 32 B_2 A \sqrt{2A_u \beta} \sigma_{\min}^{-1} r_M(\varphi)$$

then for all $n \geq n_0(A_-, A_+, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$, it holds

$$\mathbb{P}\left[\max_{k \in \{1,\dots,D\}} \left| \frac{\sqrt{rC}(P_n - P)(\psi_{1,M} \cdot \varphi_k)}{\sqrt{\sum_{j=1}^{D} (P_n - P)^2 (\psi_{1,M} \cdot \varphi_j)}} \right| \geq \frac{\tilde{R}_{n,D,\alpha}}{r_M(\varphi)\sqrt{D}} \right] \leq \frac{2D+1}{n^\beta} .$$

Proof of Lemma 3.9. By Cauchy-Schwarz inequality, we get

$$\chi_M = \sqrt{\sum_{k=1}^{D} (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} = \sup_{s \in S_M} |(P_n - P)(\psi_{1,M} \cdot s)| ,$$

where S_M is the unit sphere of M, that is

$$S_M = \left\{ s \in M,\ s = \sum_{k=1}^{D} \beta_k \varphi_k \text{ and } \sqrt{\sum_{k=1}^{D} \beta_k^2} = 1 \right\} .$$

Thus we can apply Klein-Rio's inequality (7.50) to χ_M by taking $\mathcal{F} = S_M$ and use the fact that

$$\sup_{s \in S_M} \left\| \psi_{1,M} \cdot s - P\left(\psi_{1,M} \cdot s\right) \right\|_\infty \leq 4A\sqrt{D} r_M(\varphi) \quad \text{by (3.101), (3.102) and (H4).} \tag{3.169}$$

$$\sup_{s \in S_M} \mathrm{Var}\left(\psi_{1,M} \cdot s\right) = \sup_{s \in S_M} P\left(\psi_{1,M} \cdot s\right)^2 \leq 16A^2 \quad \text{by (3.101), (3.102)}$$

and also, by using (3.169) in Inequality (7.45) applied to χ_M, we get that

$$\begin{aligned} \mathbb{E}\left[\chi_M\right] &\geq B_2^{-1} \sqrt{\mathbb{E}\left[\chi_M^2\right]} - \frac{4A\sqrt{D} r_M(\varphi)}{n} \\ &= B_2^{-1} \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - \frac{4A\sqrt{D} r_M(\varphi)}{n} . \end{aligned}$$

We thus obtain by (7.50), for all $\varepsilon, x > 0$,

$$\mathbb{P}\left(\chi_M \leq (1-\varepsilon) B_2^{-1} \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - \sqrt{32A^2 \frac{x}{n}} - \left(1 - \varepsilon + \left(1 + \frac{1}{\varepsilon}\right) x\right) \frac{4A\sqrt{D} r_M(\varphi)}{n} \right) \leq \exp(-x) . \tag{3.170}$$

So, by taking $\varepsilon = \frac{1}{2}$ and $x = \beta \ln n$ in (3.170), and by observing that $D \geq A_- (\ln n)^2$ and $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$, we conclude that, for all $n \geq n_0(A_-, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\mathbb{P}\left[\chi_M \leq \frac{B_2^{-1}}{8} \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} \right] \leq n^{-\beta} . \tag{3.171}$$

Furthermore, combining Bernstein's inequality (7.46), with the observation that we have, for every $k \in \{1, ..., D\}$,

$$\begin{aligned} \left\| \psi_{1,M} \cdot \varphi_k \right\|_\infty &\leq 4A\sqrt{D} r_M(\varphi) && \text{by (3.102) and (H4)} \\ P\left(\psi_{1,M} \cdot \varphi_k\right)^2 &\leq \left\| \psi_{1,M} \right\|_\infty^2 \leq 16A^2 && \text{by (3.102)} \end{aligned}$$

we get that, for every $x > 0$ and every $k \in \{1, ..., D\}$,

$$\mathbb{P}\left[\left| (P_n - P)\left(\psi_{1,M} \cdot \varphi_k\right) \right| \geq \sqrt{32A^2 \frac{x}{n}} + \frac{4A\sqrt{D} r_M(\varphi)}{3} \frac{x}{n} \right] \leq 2\exp(-x)$$

and so

$$\mathbb{P}\left[\max_{k \in \{1,...,D\}} \left| (P_n - P)\left(\psi_{1,M} \cdot \varphi_k\right) \right| \geq \sqrt{32A^2 \frac{x}{n}} + \frac{4A\sqrt{D} r_M(\varphi)}{3} \frac{x}{n} \right] \leq 2D\exp(-x) . \tag{3.172}$$

Hence, taking $x = \beta \ln n$ in (3.172), it comes

$$\mathbb{P}\left[\max_{k \in \{1,...,D\}} \left| (P_n - P)\left(\psi_{1,M} \cdot \varphi_k\right) \right| \geq \sqrt{\frac{32A^2 \beta \ln n}{n}} + \frac{4A\sqrt{D} r_M(\varphi) \beta \ln n}{3n} \right] \leq \frac{2D}{n^\beta} , \tag{3.173}$$

then, by using (3.171) and (3.173), we get for all $n \geq n_0(A_-, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\mathbb{P}\left[\max_{k \in \{1,...,D\}} \left| \frac{\sqrt{rC}\,(P_n - P)\left(\psi_{1,M} \cdot \varphi_k\right)}{\chi_M} \right| \geq \frac{8B_2\sqrt{rC}}{\sqrt{\frac{D}{n}} \mathcal{K}_{1,M}} \left(\sqrt{\frac{32A^2 \beta \ln n}{n}} + \frac{4A\sqrt{D} r_M(\varphi) \beta \ln n}{3n} \right) \right] \leq \frac{2D+1}{n^\beta} .$$

Finally, as $A_+ \frac{n}{(\ln n)^2} \geq D$ we have, for all $n \geq n_0 (A, A_+, r_M(\varphi), \beta)$,

$$\frac{4A\sqrt{D} r_M(\varphi) \beta \ln n}{3n} \leq \sqrt{\frac{32A^2 \beta \ln n}{n}}$$

and we can check that, since $rC \leq A_u \frac{D}{n}$ and $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$, if

$$A_\infty \geq 32 B_2 \sqrt{2A_u A^2 \beta} \sigma_{\min}^{-1} r_M(\varphi)$$

then, for all $n \geq n_0 (A_-, A_+, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\mathbb{P}\left[\max_{k \in \{1,\dots,D\}} \left| \frac{\sqrt{rC}(P_n - P)(\psi_{1,M} \cdot \varphi_k)}{\chi M} \right| \geq \frac{A_\infty}{r_M(\varphi)} \sqrt{\frac{\ln n}{n}} \right] \leq \frac{2D+1}{n^\beta}$$

which readily gives the result. ■

We are now ready to prove the lower bound (3.168) for the expected value of the largest increment of the empirical process over $\mathcal{F}_{(C,rC]}$.

Proof of Lemma 3.8. Let us begin with the lower bound of

$$\mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 ,$$

a result that will be need further in the proof. Introduce for all $k \in \{1, ..., D\}$,

$$\beta_{k,n} = \frac{\sqrt{rC}(P_n - P)(\psi_{1,M} \cdot \varphi_k)}{\sqrt{\sum_{j=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_j)}} ,$$

and observe that the excess risk on M of $\left(\sum_{k=1}^D \beta_{k,n} \varphi_k + s_M \right) \in M$ is equal to rC. We also set

$$\tilde{\Omega} = \left\{ \max_{k \in \{1,\dots,D\}} |\beta_{k,n}| \leq \frac{\tilde{R}_{n,D,\alpha}}{r_M(\varphi)\sqrt{D}} \right\} .$$

By Lemma 3.9 we have for all $\beta > 0$, if $A_\infty \geq 32 B_2 \sqrt{2A_u A^2 \beta} \sigma_{\min}^{-1} r_M(\varphi)$ then, for all $n \geq n_0 (A_-, A_+, A, B_2, r_M(\varphi), \sigma_{\min}, \beta)$,

$$\mathbb{P}\left(\tilde{\Omega}\right) \geq 1 - \frac{2D+1}{n^\beta} . \tag{3.174}$$

Moreover, by (**H4**), we get on the event $\tilde{\Omega}$,

$$\left\| \sum_{k=1}^D \beta_{k,n} \varphi_k \right\|_\infty \leq \tilde{R}_{n,D,\alpha} ,$$

and so, on $\tilde{\Omega}$,

$$\left(s_M + \sum_{k=1}^D \beta_{k,n} \varphi_k \right) \in \mathcal{F}_{(C,rC]} . \tag{3.175}$$

As a consequence, by (3.175) it holds

$$\begin{aligned}
&\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2\\
&\geq\mathbb{E}^{\frac{1}{2}}\left[\left(\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(\sum_{k=1}^{D}\beta_{k,n}\varphi_k\right)\right)\right)^2\mathbf{1}_{\tilde{\Omega}}\right]\\
&=\sqrt{rC}\sqrt{\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)\mathbf{1}_{\tilde{\Omega}}\right]}\ . \qquad (3.176)
\end{aligned}$$

Furthermore, since by (3.101) $P\left(\psi_{1,M}\cdot\varphi_k\right)=0$ and by (**H4**) $\left\|\varphi_k\right\|_\infty\leq\sqrt{D}r_M\left(\varphi\right)$ for all $k\in\{1,...,D\}$, we have

$$\begin{aligned}
\left|\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right|&\leq D\max_{k=1,...,D}\left|\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right|\\
&=D\max_{k=1,...,D}\left|P_n^2\left(\psi_{1,M}\cdot\varphi_k\right)\right|\\
&\leq D\max_{k=1,...,D}\left\|\psi_{1,M}\cdot\varphi_k\right\|_\infty^2\\
&\leq 16A^2D^2r_M^2\left(\varphi\right)
\end{aligned}$$

and it ensures

$$\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)1_{\tilde{\Omega}}\right]\geq\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)\right]-16A^2D^2r_M^2\left(\varphi\right)\mathbb{P}\left[\left(\tilde{\Omega}\right)^c\right]\ . \qquad (3.177)$$

Comparing inequality (3.177) with (3.176) and using (3.174), we obtain the following lower bound for all $n\geq n_0\left(A_-,A_+,A,B_2,r_M\left(\varphi\right),\sigma_{\min},\beta\right)$,

$$\begin{aligned}
\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2&\geq\sqrt{rC}\sqrt{\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)\right]}\\
&\quad-4Ar_M\left(\varphi\right)D\sqrt{rC}\sqrt{\mathbb{P}\left[\left(\tilde{\Omega}\right)^c\right]}\\
&\geq\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}-4Ar_M\left(\varphi\right)D\sqrt{rC}\sqrt{\frac{2D+1}{n^\beta}}\ . \qquad (3.178)
\end{aligned}$$

We take $\beta=4$, and we must have

$$A_\infty\geq 64AB_2\sqrt{2A_u}\sigma_{\min}^{-1}r_M\left(\varphi\right)\ .$$

Since $D\leq A_+n\left(\ln n\right)^{-2}$ and $\mathcal{K}_{1,M}\geq 2\sigma_{\min}$ under (**H2**), we get, for all $n\geq n_0\left(A,A_+,r_M\left(\varphi\right),\sigma_{\min}\right)$,

$$4Ar_M\left(\varphi\right)D\sqrt{rC}\sqrt{\frac{2D+1}{n^\beta}}\leq\frac{1}{\sqrt{D}}\times\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \qquad (3.179)$$

and so, by combining (3.178) and (3.179), for all $n\geq n_0\left(A_-,A_+,A,B_2,r_M\left(\varphi\right),\sigma_{\min}\right)$, it holds

$$\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2\geq\left(1-\frac{1}{\sqrt{D}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\ . \qquad (3.180)$$

Now, as $D \geq A_- (\ln n)^2$ we have for all $n \geq n_0(A_-)$, $D^{-1/2} \leq 1/2$. Moreover, we have $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ by (**H2**) and $rC \geq A_l D n^{-1}$, so we finally deduce from (3.180) that, for all $n \geq n_0(A_-, A_+, A, B_2, A_l, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 \geq \sigma_{\min} \sqrt{A_l} \frac{D}{n} \ . \tag{3.181}$$

We turn now to the lower bound of $\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s))\right]$. First observe that $s \in \mathcal{F}_{(C,rC]}$ implies that $(2s_M - s) \in \mathcal{F}_{(C,rC]}$, so that

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right] = \mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} |(P_n - P)(\psi_{1,M} \cdot (s_M - s))| \right] . \tag{3.182}$$

In the next step, we apply Corollary 7.2. More precisely, using notations of Corollary 7.2, we set

$$\mathcal{F} = \left\{ \psi_{1,M} \cdot (s_M - s) \ ; \ s \in \mathcal{F}_{(C,rC]} \right\}$$

and

$$Z = \sup_{s \in \mathcal{F}_{(C,rC]}} |(P_n - P)(\psi_{1,M} \cdot (s_M - s))| \ .$$

Now, since for all $n \geq n_0(A_+, A_-, A_\infty, A_{cons})$ we have $\tilde{R}_{n,D,\alpha} \leq 1$, we get by (3.101) and (3.102), for all $n \geq n_0(A_+, A_-, A_\infty, A_{cons})$,

$$\sup_{f \in \mathcal{F}} \|f - Pf\|_\infty = \sup_{s \in \mathcal{F}_{(C,rC]}} \|\psi_{1,M} \cdot (s_M - s)\|_\infty \leq 4A\tilde{R}_{n,D,\alpha} \leq 4A$$

we set $b = 4A$. Since we assume that $rC \leq A_u \frac{D}{n}$, it moreover holds by (3.102),

$$\sup_{f \in \mathcal{F}} \mathrm{Var}(f) \leq \sup_{s \in \mathcal{F}_{(C,rC]}} P(\psi_{1,M} \cdot (s_M - s))^2 \leq 16A^2 rC \leq 16A^2 A_u \frac{D}{n}$$

and so we set $\sigma^2 = 16A^2 A_u \frac{D}{n}$. Now, by (3.181) we have, for all $n \geq n_0(A_-, A_+, A, B_2, A_l, r_M(\varphi), \sigma_{\min})$,

$$\sqrt{\mathbb{E}[Z^2]} \geq \sigma_{\min} \sqrt{A_l} \frac{D}{n} \ . \tag{3.183}$$

Hence, a positive constant $L_{A,A_l,A_u,\sigma_{\min}}$ ($\max\left(4A\sqrt{A_u}A_l^{-1/2}\sigma_{\min}^{-1} \ ; \ 2\sqrt{A}A_l^{-1/4}\sigma_{\min}^{-1/2}\right)$ holds) exists such that, by setting

$$\varkappa_n = \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}$$

we get, using (3.183), that, for all $n \geq n_0(A_-, A_+, A_l, A_u, A, B_2, r_M(\varphi), A_{cons}, \sigma_{\min})$,

$$\varkappa_n^2 \mathbb{E}\left[Z^2\right] \geq \frac{\sigma^2}{n} \ ,$$

$$\varkappa_n^2 \sqrt{\mathbb{E}[Z^2]} \geq \frac{b}{n} \ .$$

Furthermore, since $D \geq A_- (\ln n)^2$, we have for all $n \geq n_0(A_-, A, A_u, A_l, \sigma_{\min})$,

$$\varkappa_n \in (0,1) \ .$$

So, using (3.182) and Corollary 7.2, it holds for all $n \geq n_0\left(A_-, A_+, A_l, A_u, A, B_2, r_M(\varphi), \sigma_{\min}\right)$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right] \\
&\geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}\right)\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right)^2 . \qquad (3.184)
\end{aligned}$$

Finally, by comparing (3.180) and (3.184), we deduce that for all $n \geq n_0\left(A_-, A_+, A_l, A_u, A, B_2, r_M(\varphi), \sigma_{\min}\right)$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right] \geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}$$

and so (3.168) is proved. ■

Let us now turn to the control of second order terms appearing in the expansion of the least-squares contrast, see (3.6). Let us define

$$\Omega_C(x) = \sup_{s\in\mathcal{F}_{(C,rC]}}\left\{\frac{\left|\psi_2\left((s-s_M)(x)\right) - \psi_2\left((t-s_M)(x)\right)\right|}{|s(x) - t(x)|}\ ;\ (s,t)\in\mathcal{F}_C\ ,\ s(x)\neq t(x)\right\}.$$

After straightforward computations using that $\psi_2(t) = t^2$ for all $t\in\mathbb{R}$ and assuming (**H3**), we get that, for all $x\in\mathcal{X}$,

$$\begin{aligned}
\Omega_C(x) &= 2\sup_{s\in\mathcal{F}_C}\{|s(x) - s_M(x)|\} \qquad (3.185)\\
&\leq 2\left(\tilde{R}_{n,D,\alpha}\wedge\sqrt{CD}A_{3,M}\right) . \qquad (3.186)
\end{aligned}$$

Lemma 3.10 *Let $C\geq 0$. Under (**H3**), it holds*

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|(P_n - P)\left(\psi_2\circ(s - s_M)\right)\right|\right] \leq 8\sqrt{\frac{CD}{n}}\left(\tilde{R}_{n,D,\alpha}\wedge\sqrt{CD}A_{3,M}\right) .$$

Proof. We define the Rademacher process $\mathcal{R}_n$ on a class $\mathcal{F}$ of measurable functions from $\mathcal{X}$ to $\mathbb{R}$, to be

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\ ,\quad f\in\mathcal{F}$$

where ε_i are independent Rademacher random variables also independent from the X_i. By the usual symmetrization argument we have

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|(P_n - P)\left(\psi_2\circ(s - s_M)\right)\right|\right] \leq 2\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2\circ(s - s_M)\right)\right|\right].$$

Taking the expectation with respect to the Rademacher variables, we get

$$\begin{aligned}
&\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2\circ(s - s_M)\right)\right|\right] \\
&= \mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left((s - s_M)^2\right)\right|\right] \\
&\leq \left(\max_{1\leq i\leq n}\Omega_C(X_i)\right)\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\varphi_i\left((s - s_M)(X_i)\right)\right|\right] \qquad (3.187)
\end{aligned}$$

where the functions $\varphi_i : \mathbb{R} \longrightarrow \mathbb{R}$ are defined by

$$\varphi_i(t) = \begin{cases} (\Omega_C(X_i))^{-1} t^2 & \text{for } |t| \leq \sup_{s \in \mathcal{F}_C} \{|s(X_i) - s_M(X_i)|\} = \frac{\Omega_C(X_i)}{2} \\ \frac{1}{4}\Omega_C(X_i) & \text{otherwise} \end{cases}$$

Then by (3.185) we deduce that φ_i is a contraction mapping with $\varphi_i(0) = 0$. We thus apply Theorem 7.4 to get

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_i ((s - s_M)(X_i)) \right| \right] &\leq 2\mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (s - s_M)(X_i) \right| \right] \\ &= 2\mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right] \end{aligned} \tag{3.188}$$

and so we derive successively the following upper bounds in mean,

$$\begin{aligned} &\mathbb{E}\left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (s - s_M))| \right] = \mathbb{E}\left[\mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (s - s_M))| \right] \right] \\ &\leq \mathbb{E}\left[\left(\max_{1 \leq i \leq n} \Omega_C(X_i) \right) \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_i ((s - s_M)(X_i)) \right| \right] \right] \quad \text{by (3.187)} \\ &\leq 2\mathbb{E}\left[\left(\max_{1 \leq i \leq n} \Omega_C(X_i) \right) \mathbb{E}_\varepsilon \left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right] \right] \quad \text{by (3.188)} \\ &= 2\mathbb{E}\left[\left(\max_{1 \leq i \leq n} \Omega_C(X_i) \right) \sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right] \\ &\leq 2\sqrt{\mathbb{E}\left[\max_{1 \leq i \leq n} \Omega_C^2(X_i) \right]} \sqrt{\mathbb{E}\left[\left(\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right)^2 \right]} \end{aligned}$$

We consider now an orthonormal basis of $(M, \|\cdot\|_2)$ and denote it by $(\varphi_k)_{k=1}^D$. Whence

$$\begin{aligned} &\sqrt{\mathbb{E}\left[\left(\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(s - s_M)| \right)^2 \right]} \\ &\leq \sqrt{\mathbb{E}\left[\left(\sup \left\{ \left| \sum_{k=1}^D a_k \mathcal{R}_n(\varphi_k) \right| ; \sum_{k=1}^D a_k^2 \leq C \right\} \right)^2 \right]} \\ &= \sqrt{C} \sqrt{\mathbb{E}\left[\sum_{k=1}^D (\mathcal{R}_n(\varphi_k))^2 \right]} = \sqrt{\frac{CD}{n}} \ , \end{aligned}$$

to complete the proof, it remains to observe that, by (3.186),

$$\sqrt{\mathbb{E}\left[\max_{1 \leq i \leq n} \Omega_C^2(X_i) \right]} \leq 2\left(\tilde{R}_{n,D,\alpha} \wedge \sqrt{CD} A_{3,M} \right) \ .$$

■

In the following Lemma, we provide uniform upper bounds for the supremum of the empirical process of second order terms in the contrast expansion when the considered slices are not too small.

Lemma 3.11 *Let $A_+, A_-, A_l, \beta, C_- > 0$, and assume (**H3**) and (3.17). If $C_- \geq A_l \frac{D}{n}$ and $A_+ n (\ln n)^{-2} \geq D \geq A_- (\ln n)^2$, then a positive constant $L_{A_-,A_l,\beta}$ exists such that, for all $n \geq n_0 (A_\infty, A_{cons}, A_+, A_l)$,*

$$\mathbb{P}\left[\forall C > C_-, \quad \sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_-,A_l,\beta} \sqrt{\frac{CD}{n}} \tilde{R}_{n,D,\alpha}\right] \geq 1 - n^{-\beta} .$$

Proof. First notice that, as $A_+ n (\ln n)^{-2} \geq D$, we have by (3.17),

$$\tilde{R}_{n,D,\alpha} \leq \frac{\max\left\{A_{cons} \; ; \; A_\infty \sqrt{A_+}\right\}}{\sqrt{\ln n}} .$$

By consequence, for all $n \geq n_0 (A_\infty, A_{cons}, A_+)$,

$$\tilde{R}_{n,D,\alpha} \leq 1 . \tag{3.189}$$

Now, since $\cup_{C>C_-} \mathcal{F}_C \subset B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right)$ where

$$B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right) = \left\{s \in M, \|s - s_M\|_\infty \leq \tilde{R}_{n,D,\alpha}\right\} ,$$

we have by (3.189), for all $s \in \cup_{C>C_-} \mathcal{F}_C$ and for all $n \geq n_0 (A_\infty, A_{cons}, A_+)$,

$$\begin{aligned} P(Ks - Ks_M) &= P\left[(s - s_M)^2\right] \\ &\leq \|s - s_M\|_\infty^2 \\ &\leq \tilde{R}_{n,D,\alpha}^2 \leq 1. \end{aligned}$$

We thus have, for all $n \geq n_0 (A_\infty, A_{cons}, A_+)$,

$$\bigcup_{C>C_-} \mathcal{F}_C = \bigcup_{C_- \wedge 1 < C \leq 1} \mathcal{F}_C$$

and by monotonicity of the collection $\mathcal{F}_C$, for some $q > 1$ and $J = \left\lfloor \frac{|\ln(C_- \wedge 1)|}{\ln q} \right\rfloor + 1$, it holds

$$\bigcup_{C_- \wedge 1 < C \leq 1} \mathcal{F}_C \subset \bigcup_{j=0}^{J} \mathcal{F}_{q^j C_-}.$$

Simple computations show that, since $D \geq 1$ and $C_- \geq A_l \frac{D}{n} \geq \frac{A_l}{n}$, one can find a constant $L_{A_l,q}$ such that

$$J \leq L_{A_l,q} \ln n.$$

Moreover, by monotonicity of $C \longmapsto \sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))|$, we have uniformly in $C \in \left(q^{j-1} C_-, q^j C_-\right]$,

$$\sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq \sup_{s \in \mathcal{F}_{q^{j+1} C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| .$$

Hence, taking the convention $\sup_{s \in \varnothing} |(P_n - P)(\psi_2 \circ (s - s_M))| = 0$, we get for all $n \geq n_0 (A_\infty, A_{cons}, A_+)$ and any $L > 0$,

$$\begin{aligned} &\mathbb{P}\left[\forall C > C_-, \quad \sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L \sqrt{\frac{CD}{n}} \tilde{R}_{n,D,\alpha}\right] \\ &\geq \mathbb{P}\left[\forall j \in \{1, ..., J\}, \quad \sup_{s \in \mathcal{F}_{q^j C_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L \sqrt{\frac{q^j C_- D}{n}} \tilde{R}_{n,D,\alpha}\right] . \end{aligned}$$

Now, for any $L > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\forall j \in \{1,...,J\}, \quad \sup_{s\in\mathcal{F}_{q^jC_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L\sqrt{\frac{q^jC_-D}{n}}\tilde{R}_{n,D,\alpha}\right] \\
&= 1 - \mathbb{P}\left[\exists j \in \{1,...,J\}, \quad \sup_{s\in\mathcal{F}_{q^jC_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| > L\sqrt{\frac{q^jC_-D}{n}}\tilde{R}_{n,D,\alpha}\right] \\
&\geq 1 - \sum_{j=1}^{J} \mathbb{P}\left[\sup_{s\in\mathcal{F}_{q^jC_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| > L\sqrt{\frac{q^jC_-D}{n}}\tilde{R}_{n,D,\alpha}\right] . \qquad (3.190)
\end{aligned}$$

Given $j \in \{1,...,J\}$, Lemma 3.10 yields

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{q^jC_-}} |(P_n - P)(\psi_2 \circ (s - s_M))|\right] \leq 8\sqrt{\frac{q^jC_-D}{n}}\tilde{R}_{n,D,\alpha} ,$$

and next, we apply Bousquet's inequality (7.48) to handle the deviations around the mean. We have

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_{q^jC_-}} \|\psi_2 \circ (s - s_M) - P(\psi_2 \circ (s - s_M))\|_\infty \\
&\leq 2 \sup_{s\in\mathcal{F}_{q^jC_-}} \left\|(s - s_M)^2\right\|_\infty \leq 2\tilde{R}^2_{n,D,\alpha}
\end{aligned}$$

and, for all $s \in \mathcal{F}_{q^jC_-}$,

$$\begin{aligned}
&\mathrm{Var}(\psi_2 \circ (s - s_M)) \\
&\leq P\left[(s - s_M)^4\right] \\
&\leq \|s - s_M\|^2_\infty P\left[(s - s_M)^2\right] \\
&\leq \tilde{R}^2_{n,D,\alpha} q^j C_- .
\end{aligned}$$

It follows that, for $\varepsilon = 1$ and all $x > 0$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{q^jC_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq 16\sqrt{\frac{q^jC_-D}{n}}\tilde{R}_{n,D,\alpha} + \sqrt{\frac{2\tilde{R}^2_{n,D,\alpha}q^jC_-x}{n}} + \frac{8}{3}\frac{\tilde{R}^2_{n,D,\alpha}x}{n}\right] \leq \exp(-x) . \qquad (3.191)$$

By consequence, as $D \geq A_-(\ln n)^2$ and as $\tilde{R}_{n,D,\alpha} \leq 1$ for all $n \geq n_0(A_\infty, A_{cons}, A_+)$, taking $x = \gamma \ln n$ in (3.191) for some $\gamma > 0$, easy computations show that a positive constant $L_{A_-,A_l,\gamma}$ independent of j exists such that for all $n \geq n_0(A_\infty, A_{cons}, A_+)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{q^jC_-}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_-,A_l,\gamma}\sqrt{\frac{q^jC_-D}{n}}\tilde{R}_{n,D,\alpha}\right] \leq \frac{1}{n^\gamma} .$$

Hence, using (3.190), we get for all $n \geq n_0(A_\infty, A_{cons}, A_+)$,

$$\begin{aligned}
&\mathbb{P}\left[\forall C > C_-, \quad \sup_{s\in\mathcal{F}_C} |(P_n - P)(\psi_2 \circ (s - s_M))| \leq L_{A_-,A_l,\gamma}\sqrt{\frac{CD}{n}}\tilde{R}_{n,D,\alpha}\right] \\
&\geq 1 - \frac{J}{n^\gamma} .
\end{aligned}$$

And finally, as $J \leq L_{A_l,q} \ln n$, taking $\gamma = \beta + 1$ and $q = 2$ gives the result for all $n \geq n_0 (A_\infty, A_{cons}, A_+, A_l)$.

■

Having controlled the residual empirical process driven by the remainder terms in the expansion of the contrast, and having proved sharp bounds for the expectation of the increments of the main empirical process on the slices, it remains to combine the above lemmas in order to establish the probability estimates controlling the empirical excess risk on the slices.

Lemma 3.12 *Let $\beta, A_-, A_+, A_l, C > 0$. Assume that (**H1**), (**H2**), (**H3**) and (3.17) hold. A positive constant A_4 exists, only depending on $A, A_{3,M}, \sigma_{\min}, \beta$, such that, if*

$$A_l \frac{D}{n} \leq C \leq \frac{1}{4} (1 + A_4 \nu_n)^2 \frac{D}{n} \mathcal{K}_{1,M}^2 \quad \text{and} \quad A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2$$

where $\nu_n = \max \left\{ \sqrt{\frac{\ln n}{D}},\ \sqrt{\frac{D \ln n}{n}},\ R_{n,D,\alpha} \right\}$ is defined in (3.100), then for all $n \geq n_0 (A_\infty, A_{cons}, A_+, A_l)$,

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_C} P_n (K s_M - K s) \geq \left(1 + L_{A_\infty, A, A_{3,M}, \sigma_{\min}, A_-, A_l, \beta} \times \nu_n \right) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C \right] \leq 2 n^{-\beta} .$$

Proof. Start with

$$\begin{aligned}
\sup_{s \in \mathcal{F}_C} P_n (K s_M - K s) &= \sup_{s \in \mathcal{F}_C} \{ P_n (\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ (s - s_M)) \} \\
&= \sup_{s \in \mathcal{F}_C} \{ (P_n - P) (\psi_{1,M} \cdot (s_M - s)) - (P_n - P) (\psi_2 \circ (s - s_M)) - P (K s - K s_M) \} \\
&\leq \sup_{s \in \mathcal{F}_C} \{ (P_n - P) (\psi_{1,M} \cdot (s_M - s)) - P (K s - K s_M) \} \\
&\quad + \sup_{s \in \mathcal{F}_C} |(P_n - P) (\psi_2 \circ (s - s_M))| . \qquad (3.192)
\end{aligned}$$

Next, recall that by definition,

$$D_L = \left\{ s \in B_{(M, L_\infty)} \left(s_M, \tilde{R}_{n,D,\alpha} \right),\ P (K s - K s_M) = L \right\},$$

so we have

$$\begin{aligned}
&\sup_{s \in \mathcal{F}_C} \{ (P_n - P) (\psi_{1,M} \cdot (s_M - s)) - P (K s - K s_M) \} \\
&= \sup_{0 \leq L \leq C} \sup_{s \in D_L} \{ (P_n - P) (\psi_{1,M} \cdot (s_M - s)) - L \} \\
&\leq \sup_{0 \leq L \leq C} \left\{ \sqrt{L} \sqrt{\sum_{k=1}^{D} (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} - L \right\}
\end{aligned}$$

where the last bound follows from Cauchy-Schwarz inequality. Hence, we deduce from Lemma 3.7 that

$$\mathbb{P} \left[\sup_{s \in \mathcal{F}_C} \{ (P_n - P) (\psi_{1,M} \cdot (s_M - s)) - P (K s - K s_M) \} \geq \sup_{0 \leq L \leq C} \left\{ \sqrt{L} (1 + \tau_n) \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - L \right\} \right] \leq n^{-\beta} , \qquad (3.193)$$

where

$$\begin{aligned}\tau_n &= L_{A,A_{3,M},\sigma_{\min},\beta}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right) \\ &\leq L_{A,A_{3,M},\sigma_{\min},\beta}\left(\sqrt{\frac{\ln n}{D}} \vee \sqrt{\frac{D\ln n}{n}}\right) \\ &\leq L_{A,A_{3,M},\sigma_{\min},\beta} \times \nu_n \ . \end{aligned} \tag{3.194}$$

So, injecting (3.194) in (3.193) we have

$$\mathbb{P}\left[\begin{array}{c}\sup_{s\in\mathcal{F}_C}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)-P\left(Ks-Ks_M\right)\right\} \\ \geq \sup_{0\leq L\leq C}\left\{\sqrt{L}\left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\times\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-L\right\}\end{array}\right] \leq n^{-\beta}$$

and since we assume $C \leq \frac{1}{4}\left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\times\nu_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2$ we see that

$$\sup_{0\leq L\leq C}\left\{\sqrt{L}\left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-L\right\} = \sqrt{C}\left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\times\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-C$$

and therefore

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)-P\left(Ks-Ks_M\right)\right\} \geq \left(1+L_{A,A_{3,M},\sigma_{\min},\beta}\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M}-C\right] \leq n^{-\beta}\ . \tag{3.195}$$

Moreover, as $C \geq A_l\frac{D}{n}$, we derive from Lemma 3.11 that it holds, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, A_l\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left|(P_n-P)\left(\psi_2\circ(s-s_M)\right)\right| \geq L_{A_-,A_l,\beta}\sqrt{\frac{CD}{n}}\tilde{R}_{n,D,\alpha}\right] \leq n^{-\beta}\ . \tag{3.196}$$

Finally, noticing that

$$\begin{aligned}\tilde{R}_{n,D,\alpha} &= \max\left\{R_{n,D,\alpha}, A_\infty\sqrt{\frac{D\ln n}{n}}\right\} \\ &\leq L_{A_\infty,\sigma_{\min}}\max\left\{R_{n,D,\alpha}, \sqrt{\frac{D\ln n}{n}}\right\}\times\mathcal{K}_{1,M} \qquad \text{by (3.104)} \\ &\leq L_{A_\infty,\sigma_{\min}}\times\nu_n\times\mathcal{K}_{1,M}\ ,\end{aligned}$$

we deduce from (3.196) that, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, A_l\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left|(P_n-P)\left(\psi_2\circ(s-s_M)\right)\right| \geq L_{A_\infty,\sigma_{\min},A_-,A_l,\beta}\times\nu_n\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} \tag{3.197}$$

and the conclusion follows by making use of (3.195) and (3.197) in inequality (3.192). ■
The second deviation bound for the empirical excess risk we need to establish on the upper slice is proved in a similar way.

Lemma 3.13 *Let $\beta, A_-, A_+, C \geq 0$. Assume that **(H1)**, **(H2)**, **(H3)** and (3.17) hold. A positive constant A_5, depending on $A, A_{3,M}, A_\infty, \sigma_{\min}, A_-$ and β, exists such that, if it holds*

$$C \geq \frac{1}{4}\left(1+A_5\nu_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2 \qquad \textit{and} \qquad A_+\frac{n}{(\ln n)^2} \geq D \geq A_-\left(\ln n\right)^2$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D}},\ \sqrt{\frac{D\ln n}{n}},\ R_{n,D,\alpha}\right\}$ *is defined in (3.100), then for all* $n \geq n_0\left(A_\infty, A_{cons}, A_+\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \left(1 + A_5\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq 2n^{-\beta}\ .$$

Moreover, when we only assume $C \geq 0$, *we have for all* $n \geq n_0\left(A_\infty, A_{cons}, A_+\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \frac{1}{4}\left(1 + A_5\nu_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq 2n^{-\beta}\ . \tag{3.198}$$

Proof. First observe that

$$\begin{aligned}
\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) &= \sup_{s\in\mathcal{F}_{>C}}\left\{P_n\left(\psi_{1,M}\cdot\left(s_M - s\right) - \psi_2\circ\left(s - s_M\right)\right)\right\}\\
&= \sup_{s\in\mathcal{F}_{>C}}\left\{\left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - \left(P_n - P\right)\left(\psi_2\circ\left(s - s_M\right)\right) - P\left(Ks - Ks_M\right)\right\}\\
&= \sup_{s\in\mathcal{F}_{>C}}\left\{\left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - P\left(Ks - Ks_M\right) - \left(P_n - P\right)\left(\psi_2\circ\left(s - s_M\right)\right)\right\}\\
&= \sup_{L>C}\sup_{s\in D_L}\left\{\left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - L - \left(P_n - P\right)\left(\psi_2\circ\left(s - s_M\right)\right)\right\}\\
&\leq \sup_{L>C}\left\{\sqrt{L}\sqrt{\sum_{k=1}^{D}\left(P_n - P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)} - L + \sup_{s\in\mathcal{F}_L}\left|\left(P_n - P\right)\left(\psi_2\circ\left(s - s_M\right)\right)\right|\right\}
\end{aligned} \tag{3.199}$$

where the last bound follows from Cauchy-Schwarz inequality. Now, the end of the proof is similar to that of Lemma 3.12 and follows from the same kind of computations. Indeed, from Lemma 3.7 we deduce that

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D}\left(P_n - P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)} \geq \left(1 + L_{A,A_{3,M},\sigma_{\min},\beta}\times\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} \tag{3.200}$$

and, since

$$C \geq \frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \geq \sigma_{\min}^2\frac{D}{n}\ ,$$

we apply Lemma 3.11 with $A_l = \sigma_{\min}^2$, and deduce that, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+\right)$,

$$\mathbb{P}\left[\forall L > C,\quad \sup_{s\in\mathcal{F}_L}\left|\left(P_n - P\right)\left(\psi_{2,M}^s\cdot\left(s - s_M\right)\right)\right| \geq L_{A_\infty,\sigma_{\min},A_-,\beta}\times\nu_n\sqrt{\frac{LD}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta}\ . \tag{3.201}$$

Now using (3.200) and (3.201) in (3.199) we obtain, for all $n \geq n_0\left(A_\infty, A_{cons}, A_+\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \sup_{L>C}\left\{\left(1 + L_{A,A_{3,M},A_\infty,\sigma_{\min},A_-,\beta}\times\nu_n\right)\sqrt{\frac{LD}{n}}\mathcal{K}_{1,M} - L\right\}\right] \leq 2n^{-\beta} \tag{3.202}$$

and we set $A_5 = L_{A,A_{3,M},A_\infty,\sigma_{\min},A_-,\beta}$ where $L_{A,A_{3,M},A_\infty,\sigma_{\min},A_-,\beta}$ is the constant in (3.202). For $C \geq \frac{1}{4}\left(1 + A_5\nu_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2$ we get

$$\sup_{L>C}\left\{\sqrt{L}\left(1 + A_5\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M} - L\right\} = \left(1 + A_5\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C$$

and by consequence,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \left(1+A_5\nu_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq 2n^{-\beta}\ ,$$

which gives the first part of the lemma. The second part comes from (3.202) and the fact that, for any value of $C \geq 0$,

$$\sup_{L>C}\left\{\sqrt{L}\left(1+A_5\nu_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M} - L\right\} \leq \left(1+A_5\nu_n\right)^2\frac{D}{4n}\mathcal{K}_{1,M}^2\ .$$

■

Lemma 3.14 *Let $r>1$ and $C,\beta>0$. Assume that (**H1**), (**H2**), (**H4**) and (3.17) hold and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $(M,\|\cdot\|_2)$ satisfying (**H4**). If positive constants A_-, A_+, A_l, A_u exist such that*

$$A_+\frac{n}{(\ln n)^2} \geq D \geq A_-(\ln n)^2 \quad \text{and} \quad A_l\frac{D}{n} \leq rC \leq A_u\frac{D}{n}\ ,$$

and if the constant A_∞ defined in (3.99) satisfies

$$A_\infty \geq 64B_2A\sqrt{2A_u}\sigma_{\min}^{-1}r_M(\varphi)\,,$$

then a positive constant $L_{A_-,A_l,A_u,A,A_\infty,\sigma_{\min},r_M(\varphi),\beta}$ exists such that, for all $n \geq n_0(A_-,A_+,A_u,A_l,A,A_\infty,A_{cons},B_2,r_M(\varphi),\sigma_{\min})$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_-,A_l,A_u,A,A_\infty,\sigma_{\min},r_M(\varphi),\beta}\times\nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq 2n^{-\beta}\ ,$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D}},\ \sqrt{\frac{D\ln n}{n}},\ R_{n,D,\alpha}\right\}$ is defined in (3.100).

Proof. Start with

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right)\\
&= \sup_{s\in\mathcal{F}_{(C,rC]}}\left\{\left(P_n-P\right)\left(Ks_M-Ks\right) + P\left(Ks_M-Ks\right)\right\}\\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right) - \sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_2\circ\left(s-s_M\right)\right) - \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(Ks-Ks_M\right)\\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right) - \sup_{s\in\mathcal{F}_{rC}}\left(P_n-P\right)\left(\psi_2\circ\left(s-s_M\right)\right) - rC \qquad (3.203)
\end{aligned}$$

and set

$$\begin{aligned}
S_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\\
M_{1,r,C} &= \mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right]\\
b_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}}\left\|\psi_{1,M}\cdot\left(s_M-s\right) - P\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right\|_\infty\\
\sigma_{1,r,C}^2 &= \sup_{s\in\mathcal{F}_{(C,rC]}}\mathrm{Var}\left(\psi_{1,M}\cdot\left(s_M-s\right)\right).
\end{aligned}$$

By Klein-Rio's Inequality (7.50), we get, for all $\delta, x > 0$,

$$\mathbb{P}\left(S_{1,r,C} \leq (1-\delta) M_{1,r,C} - \sqrt{\frac{2\sigma_{1,r,C}^2 x}{n}} - \left(1+\frac{1}{\delta}\right)\frac{b_{1,r,C}x}{n}\right) \leq \exp(-x) \ . \tag{3.204}$$

Then, notice that all conditions of Lemma 3.8 are satisfied, and that it gives by (3.168), for all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$M_{1,r,C} \geq \left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \ . \tag{3.205}$$

In addition, observe that

$$\sigma_{1,r,C}^2 \leq \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(\psi_{1,M}^2 \cdot (s_M - s)^2\right) \leq 16A^2rC \quad \text{by (3.102)} \tag{3.206}$$

and

$$b_{1,r,C} = \sup_{s\in\mathcal{F}_{(C,rC]}} \|\psi_{1,M}\cdot(s_M - s)\|_\infty \leq 4Ar_M(\varphi)\sqrt{rCD} \quad \text{by (3.102) and (\textbf{H4})} \tag{3.207}$$

Hence, using (3.205), (3.206) and (3.207) in inequality (3.204), we get for all $x > 0$ and all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{P}\left(S_{1,r,C} \leq (1-\delta)\left(1 - \frac{L_{A,A_l,A_u,\sigma_{\min}}}{\sqrt{D}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - \sqrt{\frac{32A^2rCx}{n}} - \left(1+\frac{1}{\delta}\right)\frac{4Ar_M(\varphi)\sqrt{rCD}x}{n}\right)$$
$$\leq \exp(-x) \ .$$

Now, taking $x = \beta\ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ and using (3.104), we deduce by simple computations that for all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1 - L_{A,A_l,A_u,\sigma_{\min},r_M(\varphi),\beta} \times \left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\right) \leq n^{-\beta} \tag{3.208}$$

and as

$$\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}} \leq \sqrt{\frac{\ln n}{D}} \vee \sqrt{\frac{D\ln n}{n}} \leq \nu_n$$

(3.208) gives, for all $n \geq n_0(A_-, A_+, A_u, A_l, A, B_2, r_M(\varphi), \sigma_{\min})$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1 - L_{A,A_l,A_u,\sigma_{\min},r_M(\varphi),\beta} \times \nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\right) \leq n^{-\beta} \ . \tag{3.209}$$

Moreover, from Lemma 3.11 we deduce that, for all $n \geq n_0(A_\infty, A_{cons}, A_+, A_l)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{rC}} |(P_n - P)(\psi_2 \circ (s - s_M))| \geq L_{A_-,A_l,\beta}\sqrt{\frac{rCD}{n}}\tilde{R}_{n,D,\alpha}\right] \leq n^{-\beta} \tag{3.210}$$

and noticing that

$$\begin{aligned}
\tilde{R}_{n,D,\alpha} &= \max\left\{R_{n,D,\alpha} \ ; \ A_\infty\sqrt{\frac{D\ln n}{n}}\right\} \\
&\leq L_{A_\infty,\sigma_{\min}} \max\left\{R_{n,D,\alpha} \ ; \ \sqrt{\frac{D\ln n}{n}}\right\} \times \mathcal{K}_{1,M} \qquad \text{by (3.104)} \\
&\leq L_{A_\infty,\sigma_{\min}} \times \nu_n \times \mathcal{K}_{1,M} \ ,
\end{aligned}$$

we deduce from (3.210) that for all $n \geq n_0\left(A_\infty, A_{cons}, A_+, A_l\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{rC}} \left|\left(P_n - P\right)\left(\psi_2 \circ \left(s - s_M\right)\right)\right| \geq L_{A_-,A_l,A_\infty,\sigma_{\min},\beta} \times \nu_n \times \sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} \ . \quad (3.211)$$

Finally, using (3.209) and (3.211) in (3.203) we get that,
for all $n \geq n_0\left(A_-, A_+, A_u, A_l, A, A_\infty, A_{cons}, B_2, r_M\left(\varphi\right), \sigma_{\min}\right)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_-,A_l,A_u,A,A_\infty,\sigma_{\min},r_M(\varphi),\beta} \times \nu_n\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq 2n^{-\beta} \ ,$$

which concludes the proof. ■

Chapitre 4

Slope heuristics in heteroscedastic bounded regression

Abstract

In this chapter, we consider the estimation of a regression function with random design and heteroscedastic noise in a non-parametric setting. More precisely, we address the problem of characterizing the optimal penalty when the regression function is estimated by using a penalized least-squares model selection method. In this context, we show the existence of a minimal penalty, defined to be the maximum level of penalization under which the model selection procedure totally misbehaves. Moreover, the optimal penalty is shown to be twice the minimal one and to satisfy a nonasymptotic pathwise oracle inequality with leading constant almost one. When the shape of the optimal penalty is known, this allows to apply the so-called *slope heuristics* initially proposed by Birgé and Massart [23], which further provides with a data-driven calibration of penalty procedure. Finally, the use of the results obtained in Chapter 3 allows us to go beyond the case of histogram models, which is already treated by Arlot and Massart in [10].

4.1 Introduction

Model selection by penalization has been the object of intensive research in the last decades. Given a collection of models and associated estimators, two different tasks can be tackled : find out the smallest true model (consistency problem), or select an estimator achieving the best performance according to some criterion, called a *risk* (efficiency problem). We only focus on the efficiency problem, where the leading idea of penalization, that goes back to early works of Akaike [1], [2] and Mallows [59], is to perform an unbiased estimation of the risk of the estimators. FPE and AIC procedures proposed by Akaike respectively in [1] and [2], as well as Mallows' C_p or C_L [59], aim to do so by adding to the empirical risk a penalty which depends on the dimension of the models. But the first analysis of such procedures had the drawback to be fundamentally asymptotic, considering in particular that the number of models as well as their dimensions are fixed while the number of data tends to infinity. As explained for example in Massart [61], various statistical situations require to let these quantities depend on the number of data. Pointing out the importance of Talagrand's type concentration inequalities in this nonasymptotic approach, Birgé and Massart [22], [26] and

Barron, Birgé and Massart [13] have thus been able to build nonasymptotic oracle inequalities for penalization procedures that take into account the complexity of the collection of models. In an abstract risk minimization framework, which includes statistical learning problems such as classification or regression, many distribution-dependent and data-dependent penalties have been proposed, from the more general and thus less accurate global penalties, see Koltchinskii [43], Bartlett & *al.* [16], to the refined local Rademacher complexities in the case where some margin relations hold (see for instance Bartlett, Bousquet and Mendelson [17], Koltchinskii [44]). But as a prize to pay for generality, the above penalties suffer from their dependence on unknown or unrealistic constants. They are very difficult to implement and calibrate in practice and satisfy oracle inequalities with possibly huge leading constants. In the general purpose, there are other penalties such as the bootstrap penalties of Efron [38] and the resampling and V-fold penalties of Arlot [7] and [5]. These penalties are essentially resampling estimates of the difference between the empirical risk and the risk and can be used in practice since, in particular, they avoid the practical drawbacks of the local Rademacher complexities. Arlot [7], [5] also proves sharp pathwise oracle inequalities for the resampling and V-fold penalties in the case of regression with random design and heteroscedastic noise on histograms models, and conjectures that the restriction on histograms is mainly technical and that his results can be extended to more general situations.

We address in this chapter the problem of optimal model selection, in a bounded heteroscedastic with random design regression setting. A penalty will be said to be optimal if it achieves a nonasymptotic oracle inequality with leading constant almost one, i.e. converging to one when the number of data tends to infinity. In the following we restrict ourselves to "small" collections of models, where the number of models is not more than polynomial in the number of data, a case where such an optimal penalty can exist. In more general settings, where the collection of models is large, one should gather the models of equal or equivalent complexity and derive an oracle inequality with respect to the infimum of the risk on the union of models with the same complexities, as explained in Birgé and Massart [23]. This would allow to consider optimal penalties for large collections of models, but this problem is anyway beyond the scope of this chapter. Birgé and Massart [23] have discovered in a generalized linear Gaussian model setting, that the optimal penalty is closely related to the minimal one, defined to be the maximal penalty under which the procedure totally misbehaves. They prove sharp upper and lower bounds for the minimal penalty and show that the optimal penalty is two times the minimal one, both for small and large collections of models. These facts are called by the authors *the slope heuristics*. The authors also exhibit a jump in the dimension of the selected model occurring around the value of the minimal penalty, and use it to estimate the minimal penalty from the data. Taking a penalty equal to two times the previous estimate then gives a nonasymptotic quasi-optimal data-driven model selection procedure. The algorithm proposed by Birgé and Massart [23] to estimate the minimal penalty relies on the previous knowledge of the shape of the latter, which is a known function of the dimension of the models in their setting, and thus their procedure gives a data-driven *calibration* of the minimal penalty. Considering the case of Gaussian least-squares regression with unknown variance, Baraud, Giraud and Huet [11] have also derived lower bounds for the penalty terms for small and large collection of models, as well as Castellan [30] in the case of maximum likelihood estimation of density on histograms where a lower bound on the penalty term is given only for small collections of models - see also Chapter 5 where we validate the slope heuristics for the maximum likelihood estimation of density on histograms and propose two directions of generalization. Then the slope phenomenon has been extended by Arlot and Massart [10] in a bounded heteroscedastic with random design regression framework. They consider least-squares estimators on a "small" collection of histograms models. Heteroscedasticity of the noise allows them to validate the slope heuristics without assuming a particular shape of the penalty, and in particular to consider situations where the shape of the penalty is not a function of the dimension of the models.

In such general cases, the authors propose to estimate the shape of the penalty by using Arlot's resampling or V-fold penalties, proved to be efficient in their regression framework by Arlot [5] and [7], in order to derive an accurate data-driven calibration of the optimal penalty. Moreover, their approach is more general than the histogram case, except for some identified technical parts of their proofs, thus providing with some quite general algebra that can be applied in other frameworks to derive sharp model selection results. The authors have also identified the minimal penalty as the mean of the empirical excess risk on each model, and the ideal penalty to be estimated as the sum of the empirical excess risk and true excess risk on each model. The slope heuristics then heavily relies on the fact that the empirical excess risk is equivalent to the true excess risk for models of reasonable dimensions. Arlot and Massart [10] conjecture that this equivalence between the empirical and true excess risk is a quite general fact in M-estimation, as well as, by rather direct consequence, the slope phenomenon for models not too badly chosen in terms of approximation properties. A general result supporting this conjecture is the high dimensional Wilks' phenomenon discovered by Boucheron and Massart [27] in the setting of bounded contrast minimization under margin conditions, where the authors derive concentrations inequalities for the true and empirical excess risk when the considered model satisfies some general condition on the moment of first order of the surpremum of the empirical process on localized slices of variance in the loss class. This assumption can be explicated under suitable covering entropy conditions on the model. Lerasle [56] proved the validity of the slope heuristics in a least-squares density estimation setting, under rather mild conditions on the considered linear models. The approach developed by Lerasle in this framework allows sharp computations and the empirical excess risk is shown by the author to be exactly equal to the true excess risk. Moreover, some improvements comparing to the technology of proofs given by Arlot and Massart [10] can be found in [56], where Lerasle considers comparison between all pairs of models, allowing in particular a more refined use of the bias of the models. Lerasle also proves in the least-squares density estimation setting the efficiency of Arlot's resampling penalties, and generalizes these results for weakly dependent data, see [57]. Arlot and Bach [8] recently consider the problem of selecting among linear estimators in non-parametric regression. Their framework includes model selection for linear regression, the choice of a regularization parameter in kernel ridge regression or spline smoothing, and the choice of a kernel in multiple kernel learning. In such cases, the minimal penalty is not necessarily half the optimal one, but the authors propose to estimate the unknown variance by the minimal penalty and to use it in a plug-in version of Mallows' C_L. The latter penalty is proved to be optimal by establishing a nonasymptotic oracle inequality with constant almost one.

In this chapter, we prove the validity of the slope heuristics in a bounded heteroscedastic with random design regression framework, by considering a "small" collection of finite-dimensional linear models, a setting that extends the case of histograms already treated by Arlot and Massart [10]. Two main assumptions must be satisfied. First, we require that the models have a uniform localized orthonormal basis structure in $L_2\left(P^X\right)$, where P^X is the law of the explicative variable X. This kind of analytical property describing the L_∞-structure of the models has already been used in a model selection framework by Birgé and Massart [22] and Barron, Birgé and Massart [13] (see also Massart [61]). Considering for example the unit cube of $\mathbb{R}^q$ and taking $P^X = \text{Leb}$ the Lebesgue measure on it, it is shown in Birgé and Massart [22] that the assumption of localized orthonormal basis are satisfied for some wavelet expansions and piecewise polynomials uniformly bounded in their degrees. It is also known, Massart [61], that in the case of histograms the property of localized basis in $L_2\left(P^X\right)$ is equivalent to the lower regularity of the considered partition with respect to P^X, an assumption required by Arlot and Massart in [10]. Moreover, we show in Chapter 3 that if P^X has a density with respect to the Lebesgue measure on the unit interval that is uniformly bounded away from zero then, assuming the lower regularity of the partition defining piecewise polynomials of uniformly bounded degrees ensures that the assumption of localized basis is satisfied for such a model.

The second property that must be satisfied in our setting is that the least-squares estimators are uniformly consistent over the collection of models and converge to the orthogonal projections of the unknown regression function. Again, such a property is shown in Chapter 3 to be satisfied for suitable histograms and more general piecewise polynomial models. This allows us to recover the results of Arlot and Massart [10] with the same set of assumptions when the noise is uniformly bounded by upper and by below, and to extend it to models of piecewise polynomials uniformly bounded in their degrees. Taking advantage of the sharp estimates of the empirical and true excess risks for a fixed model given in Chapter 3, our proofs then rely on the same algebra of proofs as those given in Arlot and Massart [10]. Moreover, our arguments rely on the more general concept of regular contrast exposed in Chapter 7, and can be extended to other frameworks than least-squares regression, see for example Chapters 5 and 6 for applications in the maximum likelihood and least-squares estimation of density, respectively.

The chapter is organized as follows. We describe in Section 4.2 the statistical framework, the slope heuristics and the subsequent data-driven algorithm of calibration of penalties. We state in Section 4.3 our main results and derive their proofs in the remainder of the chapter.

4.2 Statistical framework and the slope heuristics

4.2.1 Penalized least-squares model selection

We assume that we have n independent observations $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ with common distribution P. The marginal law of X_i is denoted by P^X. We assume that the data satisfy the following relation

$$Y_i = s_* (X_i) + \sigma (X_i) \varepsilon_i \ , \tag{4.1}$$

where $s_* \in L_2 \left(P^X\right)$, ε_i are i.i.d. random variables with mean 0 and variance 1 conditionally to X_i and $\sigma : \mathcal{X} \longrightarrow \mathbb{R}$ is an heteroscedastic noise level. A generic random variable of law P, independent of the sample $(\xi_1, ..., \xi_n)$, is denoted by $\xi = (X, Y)$.
Hence, s_* is the regression function of Y with respect to X, that we want to estimate. We are given a finite collection of models $\mathcal{M}_n$, with cardinality depending on the number of data n. Each model $M \in \mathcal{M}_n$ is assumed to be a finite-dimensional vector space, and we denote by D_M its linear dimension and s_M the linear projection of s_* onto M in $L_2 \left(P^X\right)$. Furthermore, by setting $K : L_2 \left(P^X\right) \longrightarrow L_1 (P)$ the least-squares contrast, defined by

$$K (s) = (x, y) \longmapsto (y - s (x))^2 , \quad s \in L_2 \left(P^X\right) \ ,$$

the regression function s_* satisfy

$$s_* = \arg \min_{s \in L_2 (P^X)} PK (s)$$

and for the linear projections s_M we have

$$s_M = \arg \min_{s \in M} PK (s) \ .$$

For each model $M \in \mathcal{M}_n$, we consider a least-squares estimator $s_n (M)$, satisfying

$$\begin{aligned} s_n (M) &\in \arg \min_{s \in M} \{P_n (K (s))\} \\ &= \arg \min_{s \in M} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - s (X_i))^2 \right\} \end{aligned}$$

where $P_n = n^{-1} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$ is the empirical measure built from the data. We measure the performance of the least-squares estimators by their excess risk,

$$l (s_*, s_n (M)) := P (K s_n (M) - K s_*) = \| s_n (M) - s_* \|_2^2$$

where $\|s\|_2 = \left(\int_{\mathcal{X}} s^2 dP^X\right)^{1/2}$ is the quadratic norm in $L_2\left(P^X\right)$. Moreover, we have

$$l\left(s_*, s_n\left(M\right)\right) = l\left(s_*, s_M\right) + l\left(s_M, s_n\left(M\right)\right) \ ,$$

where the quantity

$$l\left(s_*, s_M\right) := P\left(Ks_M - Ks_*\right) = \|s_M - s_*\|_2^2$$

is called the bias of the model M and $l\left(s_M, s_n\left(M\right)\right) := P\left(Ks_n\left(M\right) - Ks_M\right) \geq 0$ is the excess risk of the least-squares estimator $s_n\left(M\right)$ on M. By the Pythagorean identity, we have

$$l\left(s_M, s_n\left(M\right)\right) = \|s_n\left(M\right) - s_M\|_2^2$$

and we prove sharp bounds for the latter quantity in Chapter 3, based on the concept of regular contrast exposed in Chapter 2.
Given the collection of models $\mathcal{M}_n$, an oracle model M_* is defined to be

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \left\{l\left(s_*, s_n\left(M\right)\right)\right\} \tag{4.2}$$

and the associated oracle estimator $s_n\left(M_*\right)$ thus achieves the best performance in terms of excess risk among the collection $\left\{s_n\left(M\right); M \in \mathcal{M}_n\right\}$. Unfortunately, the oracle model is unknown as it depends on the unknown law P of the data, and we propose to estimate it by a model selection procedure via penalization. Given some known penalty pen, that is a function from $\mathcal{M}_n$ to $\mathbb{R}_+$, we thus consider the following data-dependent model, also called selected model,

$$\widehat{M} \in \arg\min_{M \in \mathcal{M}_n} \left\{P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}\left(M\right)\right\} \ . \tag{4.3}$$

Our goal is then to find a good penalty, such that the selected model $\widehat{M}$ satisfies an oracle inequality of the form

$$l\left(s_*, s_n\left(\widehat{M}\right)\right) \leq C \times \ell\left(s_*, s_n\left(M_*\right)\right) \ ,$$

with some positive constant C as close to one as possible and with high probability, typically more than $1 - Ln^{-2}$ for some positive constant L.

4.2.2 The slope heuristics

Let us rewrite the definition of the oracle model M_* given in (4.2). As for any $M \in \mathcal{M}_n$, the excess risk $l\left(s_*, s_n\left(M\right)\right) = P\left(Ks_n\left(M\right)\right) - P\left(Ks_*\right)$ is the difference between the risk of the estimator $s_n\left(M\right)$ and the risk of the target s_*, and as $P\left(Ks_*\right)$ is a constant of the problem, it holds

$$\begin{aligned} M_* &\in \arg\min_{M \in \mathcal{M}_n} \left\{P\left(Ks_n\left(M\right)\right)\right\} \\ &= \arg\min_{M \in \mathcal{M}_n} \left\{P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}_{\mathrm{id}}\left(M\right)\right\} \end{aligned}$$

where for all $M \in \mathcal{M}_n$,

$$\mathrm{pen}_{\mathrm{id}}\left(M\right) := P\left(Ks_n\left(M\right)\right) - P_n\left(Ks_n\left(M\right)\right) \ .$$

The penalty function $\mathrm{pen}_{\mathrm{id}}$ is called the *ideal penalty*, as it allows to select the oracle, but it is unknown because it depends on the distribution of the data. As pointed out by Arlot and Massart [10], the leading idea of penalization in the efficiency problem is thus to give some sharp estimate of the ideal penalty, in order to perform an unbiased or asymptotically unbiased estimation of the risk over the collection of models, leading to a sharp oracle inequality for the selected model. A penalty term $\mathrm{pen}_{\mathrm{opt}}$ is said to be optimal if it achieves an oracle inequality with constant almost one, tending to one when the number n of data tends to infinity.
Concerning the estimation of the optimal penalty, Arlot and Massart [10] conjecture that the mean of the empirical excess risk $\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$ satisfies the following slope heuristics in a quite general framework:

(i) If a penalty $\text{pen} : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ is such that, for all model $M \in \mathcal{M}_n$,

$$\text{pen}(M) \leq (1-\delta)\,\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right]$$

with $\delta > 0$, then the dimension of the selected model $\widehat{M}$ is "very large" and the excess risk of the selected estimator $s_n\left(\widehat{M}\right)$ is "much larger" than the excess risk of the oracle.

(ii) If $\text{pen} \approx (1+\delta)\,\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right]$ with $\delta > 0$, then the corresponding model selection procedure satisfies an oracle inequality with a leading constant $C(\delta) < +\infty$ and the dimension of the selected model is "not too large". Moreover,

$$\text{pen}_{\text{opt}} \approx 2\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right]$$

is an optimal penalty.

The mean of the empirical excess risk on M, when M varies in $\mathcal{M}_n$, is thus conjectured to be the maximal value of penalty under which the model selection procedure totally misbehaves. It is called the *minimal penalty,* denoted by $\text{pen}_{\min}$:

$$\text{for all } M \in \mathcal{M}_n, \qquad \text{pen}_{\min}(M) = \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] \ .$$

The optimal penalty is then close to two times the minimal one,

$$\text{pen}_{\text{opt}} \approx 2\,\text{pen}_{\min} \ .$$

Let us now briefly explain why points (i) and (ii) below are natural. We give in Section 4.3 precise results which validate the slope heuristics for models such as histograms or piecewise polynomials uniformly bounded in their degrees. If the penalty is the minimal one, then for all $M \in \mathcal{M}_n$,

$$\begin{aligned}
&P_n\left(Ks_n(M)\right) + \text{pen}_{\min}(M) \\
&= P_n\left(Ks_n(M)\right) + \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] \\
&= P\left(Ks_M\right) + \left(P_n - P\right)\left(Ks_M\right) + \left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] - P_n\left(Ks_M - Ks_n(M)\right)\right) \\
&\approx P\left(Ks_M\right) \ .
\end{aligned}$$

In the above lines, we neglect $\left(P_n - P\right)\left(Ks_M\right)$ as it is a centered quantity and if the empirical excess risk $P_n\left(Ks_n(M) - Ks_M\right)$ is close enough to its expectation, then the selected model almost minimizes its bias, and so its dimension is among the largest of the models and the excess risk of the selected estimator blows up. As shown by Boucheron and Massart [27], the empirical excess risk satisfies a concentration inequality in a general framework, which allows to neglect the difference with its mean, at least for models that are not too small.
Now, if the chosen penalty is less than the minimal one, $\text{pen} \approx (1-\delta)\,\text{pen}_{\min}$ with $\delta \in (0,1)$, the algorithm minimizes over $\mathcal{M}_n$,

$$\begin{aligned}
&P_n\left(Ks_n(M)\right) + \text{pen}(M) \\
&\approx P\left(Ks_M\right) - \delta P_n\left(Ks_M - Ks_n(M)\right) + \left(P_n - P\right)\left(Ks_M\right) \\
&\quad + (1-\delta)\left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] - P_n\left(Ks_M - Ks_n(M)\right)\right) \\
&\approx P\left(Ks_M\right) - \delta P_n\left(Ks_M - Ks_n(M)\right) \ ,
\end{aligned}$$

where in the last identity we neglect the deviations of the empirical excess risk and the difference between the empirical and true risk of the projections s_M. As the empirical excess risk is increasing and the risk of the projection s_M is decreasing with respect to the complexity of the

models, the penalized criterion is decreasing with respect to the complexity of the models, and the selected model is again among the largest of the collection.
If on the contrary, the chosen penalty is more than the minimal one, $\text{pen} \approx (1+\delta)\,\text{pen}_{\min}$ with $\delta > 0$, then the selected model minimizes the following criterion, for all $M \in \mathcal{M}_n$,

$$\begin{aligned} &P_n\left(Ks_n\left(M\right)\right) + \text{pen}\left(M\right) - P_n\left(Ks_*\right) \\ &\approx \ell\left(s_*, s_M\right) + \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) + \left(P_n - P\right)\left(Ks_M - Ks_*\right) \\ &\quad + (1+\delta)\left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] - P_n\left(Ks_M - Ks_n\left(M\right)\right)\right) \\ &\approx \ell\left(s_*, s_M\right) + \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) \ , \end{aligned} \tag{4.4}$$

So the selected model achieves a trade-off between the bias of the models which decreases with the complexity and the empirical excess risk which increases with the complexity of the models. The selected dimension will be then reasonable, and the trade-off between the bias and the complexity of the models is likely to give some oracle inequality.
Finally, if we take $\delta = 1$ in the above case, that is $\text{pen} \approx 2 \times \text{pen}_{\min}$ and if we assume that the empirical excess risk is equivalent to the excess risk,

$$P_n\left(Ks_M - Ks_n\left(M\right)\right) \sim P\left(Ks_n\left(M\right) - Ks_M\right) \ , \tag{4.5}$$

then according to (4.4) the selected model almost minimizes

$$P\left(Ks_M - Ks_*\right) + P_n\left(Ks_M - Ks_n\left(M\right)\right) \approx \ell\left(s_*, s_M\right) + P\left(Ks_n\left(M\right) - Ks_M\right) \approx \ell\left(s_*, s_n\left(M\right)\right) \ .$$

Hence,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \approx \ell\left(s_*, s_n\left(M_*\right)\right)$$

and the procedure is nearly optimal. We give in Chapter 3 some results showing that (4.5) is a quite general fact in least-squares regression.

4.2.3 A data-driven calibration of penalty algorithm

The slope heuristics stated in points (i) and (ii) in Section 4.2.2, include that a jump in the dimensions of the selected models should occur around the minimal penalty, which can be used to estimate the minimal penalty and by consequence, the optimal one. Let us denote by $\text{pen}_{\text{shape}}$ the shape of the minimal penalty which is, according to the slope heuristics, equal to the shape of the optimal penalty. Thus, for two unknown positive constants $A_{\min}$ and A_* depending on the unknown distribution of the data, we have

$$\text{pen}_{\min} = A_{\min}\,\text{pen}_{\text{shape}} \quad \text{and} \quad \text{pen}_{\text{opt}} = A_*\,\text{pen}_{\text{shape}} \ ,$$

where

$$A_* = 2 \times A_{\min}$$

whenever the optimal penalty is twice the minimal one. We assume now that the shape of the minimal penalty is known, from some prior knowledge or because it has been estimated from the data, for example by using Arlot's resampling and V-fold penalties as suggested in Arlot and Massart [10]. Then, Arlot and Massart [10] propose to *calibrate* the optimal penalty by the following procedure and by doing so, they extend to general penalty shapes a previous algorithm proposed by Birgé and Massart [23].

Algorithm of data-driven calibration of penalties :

1. Compute the selected model $\widehat{M}\left(A\right)$ as a function of $A > 0$,

$$\widehat{M}\left(A\right) \in \arg\min_{M \in \mathcal{M}_n} \left\{P_n K\left(s_n\left(M\right)\right) + A\,\text{pen}_{\text{shape}}\left(M\right)\right\} \ .$$

2. Find $\hat{A}_{\min} > 0$ such that the dimension $D_{\widehat{M}(A)}$ is "very large" for $A < \hat{A}_{\min}$ and "reasonably small" for $A > \hat{A}_{\min}$.

3. Select the model $\widehat{M} = \widehat{M}\left(2\hat{A}_{\min}\right)$.

In this chapter, since our aim is not to apply the above algorithm in practice, we refer to Arlot and Massart [10] for a detailed presentation of the algorithm and to Baudry, Maugis and Michel [21] for an overview on the slope heuristics and further discussions on implementation issues. Data-driven calibration of penalties algorithms have already been applied successively in many statistical frameworks such as mixture models [63], clustering [20], spatial statistics [86], estimation of oil reserves [54] and genomics [87], to name but a few. These applications tend to support the conjecture of Arlot and Massart [10] that the slope heuristics is valid in a quite general framework.

4.3 Main Results

We state here our results that theoretically validate the slope heuristics in our bounded heteroscedastic regression setting. In particular, we recover the results stated in Theorems 2 and 3 of Arlot and Massart [10] for histogram models and extend them to models of piecewise polynomials uniformly bounded in their degrees. The proofs are postponed to the end of the chapter, and heavily rely on results of Chapter 3 where we consider a fixed model, and on the general algebra of proofs developed by Arlot and Massart [10]. We state now the assumptions required to derive our results.

4.3.1 Main assumptions

Let us begin with the set of assumptions needed in the general case of models that are provided with localized basis in $L_2\left(P^X\right)$.

General set of assumptions : (GSA)

(P1) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}\left(\mathcal{M}_n\right) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(P2) Upper bound on dimensions of models in $\mathcal{M}_n$: there exists a positive constant $A_{\mathcal{M},+}$ such that for every $M \in \mathcal{M}_n$, $1 \leq D_M \leq A_{\mathcal{M},+} n \left(\ln n\right)^{-2} \leq n$.

(P3) Richness of $\mathcal{M}_n$: there exist $M_0, M_1 \in \mathcal{M}_n$ such that $D_{M_0} \in [\sqrt{n}, c_{rich}\sqrt{n}]$ and $D_{M_1} \geq A_{rich} n \left(\ln n\right)^{-2}$.

(Ab) A positive constant A exists, that bounds the data and the projections s_M of the target s_* over the models M of the collection $\mathcal{M}_n$: $|Y_i| \leq A < \infty$, $\|s_M\|_\infty \leq A < \infty$ for all $M \in \mathcal{M}_n$.

(An) Uniform lower-bound on the noise level: $\sigma\left(X_i\right) \geq \sigma_{\min} > 0$ *a.s.*

(Ap$_u$) The bias decreases as a power of D_M: there exist $\beta_+ > 0$ and $C_+ > 0$ such that

$$\ell\left(s_*, s_M\right) \leq C_+ D_M^{-\beta_+} .$$

(Alb) Each model is provided with a localized basis: there exists a constant $r_{\mathcal{M}}$ such that for each $M \in \mathcal{M}_n$ one can find an orthonormal basis $(\varphi_k)_{k=1}^{D_M}$ satisfying that, for all $(\beta_k)_{k=1}^{D_M} \in \mathbb{R}^{D_M}$,

$$\left\| \sum_{k=1}^{D_M} \beta_k \varphi_k \right\|_\infty \leq r_{\mathcal{M}} \sqrt{D_M} \, |\beta|_\infty \ ,$$

where $|\beta|_\infty = \max \{ |\beta_k| \ ; \ k \in \{1, ..., D_M\} \}$.

($\mathbf{Ac}_\infty$) Consistency in sup-norm of least-squares estimators: an event Ω_∞ of probability at least $1 - n^{-2-\alpha_{\mathcal{M}}}$, a positive constant A_{cons}, a positive integer n_1 and a collection of positive numbers $(R_{n,D_M})_{M \in \mathcal{M}_n}$ exist, such that

$$\sup_{M \in \mathcal{M}_n} R_{n,D_M} \leq \frac{A_{cons}}{\sqrt{\ln n}} \tag{4.6}$$

and for all $M \in \mathcal{M}_n$ it holds on Ω_∞, for all $n \geq n_1$,

$$\| s_n (M) - s_M \|_\infty \leq R_{n,D_M} \ . \tag{4.7}$$

We turn now to the set of assumptions needed for histogram models and models by piecewise polynomials, respectively.

Set of assumptions for histogram models :

Given some linear histogram model $M \in \mathcal{M}_n$, we denote by $\mathcal{P}_M$ the associated partition of $\mathcal{X}$. Take assumptions (**P1**), (**P2**), (**P3**), (**An**) and ($\mathbf{Ap}_u$) from the general set of assumptions. Assume moreover that the following conditions hold true:

(Ab') A positive constant A exists, that bounds the data: $|Y_i| \leq A < \infty$.

(Alrh) Lower regularity of the partitions: there exists a positive constant $c^h_{\mathcal{M},P}$ such that,

$$\text{for all } M \in \mathcal{M}_n, \quad \sqrt{|\mathcal{P}_M| \inf_{I \in \mathcal{P}_M} P^X (I)} \geq c^h_{\mathcal{M},P} > 0 \ .$$

Set of assumptions for piecewise polynomials models :

In this case we take $\mathcal{X} = [0,1]$, Leb is the Lebesgue measure on $\mathcal{X}$, and given a linear model $M \in \mathcal{M}_n$ of piecewise polynomials, we denote by $\mathcal{P}_M$ the associated partition of $\mathcal{X}$.
Take assumptions (**P1**), (**P2**), (**P3**), (**An**) and ($\mathbf{Ap}_u$) from the general set of assumptions. Assume moreover that the following additional conditions hold.

(Ab') A positive constant A exists, that bounds the data: $|Y_i| \leq A < \infty$.

(Aud) Uniformly bounded degrees: there exists $r \in \mathbb{N}^*$ such that, for all $M \in \mathcal{M}_n$, all $I \in \mathcal{P}_M$ and all $p \in M$,

$$\deg \left(p_{|I} \right) \leq r \ .$$

($\mathbf{Ad}_{\text{Leb}}$) Density bounded from upper and from below: P^X has a density f with respect to Leb satisfying for some constants $c_{\min}$ and $c_{\max}$, that

$$0 < c_{\min} \leq f(x) \leq c_{\max} < \infty, \quad \forall x \in [0,1] \ .$$

(Alrpp) Lower regularity of the partition: a positive constant $c^{pp}_{\mathcal{M},P}$ exists such that, for all $M \in \mathcal{M}_n$,

$$0 < c^{pp}_{\mathcal{M},\text{Leb}} \leq \sqrt{|\mathcal{P}_M| \inf_{I \in \mathcal{P}_M} \text{Leb}(I)} < +\infty \ .$$

The sets of assumptions will be discussed in Section 4.3.3.

4.3.2 Statement of the theorems

Theorem 4.1 *Under the general set of assumptions (**GSA**) of Section 4.3.1, for $A_{\text{pen}} \in [0,1)$ and $A_p > 0$, we assume that with probability at least $1 - A_p n^{-2}$ we have*

$$0 \leq \text{pen}\left(M_1\right) \leq A_{\text{pen}} \mathbb{E}\left[P_n\left(K s_M - K s_n\left(M_1\right)\right)\right] \ , \tag{4.8}$$

*where the model M_1 is defined in assumption (**P3**) of (**GSA**). Then there exist two positive constants A_1, A_2 independent of n such that, with probability at least $1 - A_1 n^{-2}$, we have, for all $n \geq n_0\left(\left(\boldsymbol{GSA}\right), A_{\text{pen}}\right)$,*

$$D_{\widehat{M}} \geq A_2 n \ln\left(n\right)^{-2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \ln\left(n\right) \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{4.9}$$

Moreover, in the case of histograms and piecewise polynomials models, taking their respective set of assumptions defined in Section 4.3.1 yields the same results.

Thus, Theorem 4.1 justifies the first part (**i**) of the slope heuristics exposed in Section 4.2.2. As a matter of fact, it shows that there exists a level such that if the penalty is smaller than this level for one of the largest models, then the dimension of the output is among the largest dimensions of the collection and the excess risk of the selected estimator is much bigger than the excess risk of the oracle. Moreover, this level is given by the mean of the empirical excess risk of the least-squares estimator on each model.
The following theorem validates the second part of the slope heuristics.

Theorem 4.2 *Assume that the general set of assumptions (**GSA**) of Section 4.3.1 hold. Moreover, for some $\delta \in [0,1)$ and $A_p, A_r > 0$, assume that an event of probability at least $1 - A_p n^{-2}$ exists on which, for every model $M \in \mathcal{M}_n$ such that $D_M \geq A_{\mathcal{M},+}\left(\ln n\right)^3$, it holds*

$$(2-\delta)\,\mathbb{E}\left[P_n\left(K s_M - K s_n\left(M\right)\right)\right] \leq \text{pen}\left(M\right) \leq (2+\delta)\,\mathbb{E}\left[P_n\left(K s_M - K s_n\left(M\right)\right)\right] \tag{4.10}$$

together with

$$\text{pen}\left(M\right) \leq A_r \frac{\left(\ln n\right)^3}{n} \tag{4.11}$$

*for every model $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+}\left(\ln n\right)^3$. Then, for $\frac{1}{2} > \eta > \left(1 - \beta_+\right)_+ / 2$, there exist a positive constant A_3 only depending on $c_{\mathcal{M}}$ given in (**GSA**) and on A_p, a positive constant A_4 only depending on constants in the set of assumptions (**GSA**), a positive constant A_5 only depending on constants in the set of assumptions (**GSA**) and on A_r and a sequence*

$$\theta_n = A_4 \sup_{M \in \mathcal{M}_n} \left\{\varepsilon_n\left(M\right), \ A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_M \leq n^{\eta+1/2}\right\} \leq \frac{A_4\left(1 \vee \sqrt{A_{cons}}\right)}{\left(\ln n\right)^{1/4}} \tag{4.12}$$

such that with probability at least $1 - A_3 n^{-2}$, it holds for all $n \geq n_0\left(\left(\boldsymbol{GSA}\right), \eta, \delta\right)$,

$$D_{\widehat{M}} \leq n^{\eta+1/2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\left(\left(\ln n\right)^{-2} + \theta_n\right)}{\left(1-\delta\right)^2}\right) \ell\left(s_*, s_n\left(M_*\right)\right) + A_5 \frac{\left(\ln n\right)^3}{n} \ . \tag{4.13}$$

Assume that in addition, the following assumption holds,

(Ap) *The bias decreases like a power of D_M : there exist $\beta_- \geq \beta_+ > 0$ and $C_+, C_- > 0$ such that*

$$C_- D_M^{-\beta_-} \leq \ell\left(s_*, s_M\right) \leq C_+ D_M^{-\beta_+} \ .$$

Then it holds with probability at least $1 - A_3 n^{-2}$, for all $n \geq n_0\left(\left(\boldsymbol{GSA}\right), C_-, \beta_-, \eta, \delta\right)$,

$$A_{\mathcal{M},+} \left(\ln n\right)^3 \leq D_{\widehat{M}} \leq n^{\eta+1/2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{(1-\delta)^2}\right) \ell\left(s_*, s_n\left(M_*\right)\right) \ . \tag{4.14}$$

*Likewise, in the case of models of histograms and piecewise polynomials, taking their respective set of assumptions defined in Section 4.3.1, together with assumption (4.10) and, for the second part of the theorem, assumption (**Ap**), yields the same results.*

The quantity $\varepsilon_n\left(M\right)$ used in (4.12) controls the deviations of the true and empirical excess risks on the model M and is more precisely defined in Remark 4.1 above. From Theorems 4.1 and 4.2, we identify the minimal penalty with the mean of the empirical excess risk on each model,

$$\mathrm{pen}_{\min}\left(M\right) = \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \ .$$

Moreover, Theorem 4.2 states in particular that if the penalty is close to two times the minimal procedure, then the selected estimator satisfies a pathwise oracle inequality with constant almost one, and so the model selection procedure is approximately optimal.

4.3.3 Comments on the sets of assumptions

Let us now explain the sets of assumptions given in Section 4.3.1. Assumption (**P1**) states that the collection of models has a small complexity, more precisely a polynomially increasing one with respect to the amount of data. For this kind of complexities, if one wants to perform a good model selection procedure for prediction, the chosen penalty should estimate the mean of the ideal one on each model. Indeed, as Talagrand's type inequalities for the empirical process are pre-Gaussian, they allow to neglect the deviations of the quantities of interest from their mean, uniformly over the collection of models. This is not the case for too large collections of models, where one has to put an extra-log factor depending on the complexity of the collection of models inside the penalty (see for example [22] and [13]). In assumption (**P2**) we restrict the dimensions of the models by upper, in a way that is not too restrictive since we allow the dimension to be of the order of the amount of data within a power of a logarithmic factor. We assume in (**P3**) that the collection of models contains a model M_0 of reasonably large dimension and a model M_1 of high dimension, which is necessary since we prove the existence of a jump between high and reasonably large dimensions. We demand in ($\mathbf{Ap}_u$) that the quality of approximation of the collection of models is good enough in terms of bias. More precisely, we require a polynomially decreasing of excess risk of linear projections of the regression function onto the models. Assumptions (**Ab**), (**An**), (**Alb**) and ($\mathbf{Ac}_\infty$) essentially allow us to apply results of Section 3.3, as further explained in Remark 4.1 below. The assumption (**Ab**) is also necessary to control in the proofs the empirical bias term centered by the true bias by using Bernstein's inequality (see Lemma 4.2).
Assumption (**Ab'**) implies in the histogram case assumption (**Ab**), see Section 3.4.4 of Chapter 3. Moreover, assumption (**Alrh**) allows us in this case to deduce assumptions (**Alb**) and ($\mathbf{Ac}_\infty$) of the general set of assumptions (see Lemma 3.2 and 3.3 of Chapter 3). Moreover, using Lemma 3.3, it is straightforward to see that in the histogram case we have

$$R_{n,D_M} \leq A_{cons} \sqrt{\frac{D_M \ln n}{n}} \ ,$$

where A_{cons} is a uniform positive constant over the models of $\mathcal{M}_n$. We obtain in the case of histograms the same set of assumptions as given in Arlot and Massart [10]. Arlot and Massart [10] also notice that they can weaken assumptions (**Ab'**) and (**An**), for example by assuming conditions on the moment of the noise instead of considering that this quantity is bounded in sup-norm. This latter improvement seems to be beyond the reach of our method, due to the use of Talagrand's type inequalities that require conditions in sup-norm. Arlot and Massart [10] also show that the condition ($\mathbf{Ap}_u$) is satisfied when $\mathcal{X} \subset \mathbb{R}^k$ and the regression function s_* is α-Hölderian. Moreover, they show that (**Ap**) is satisfied when in addition, s_* is non-constant with respect to the sup-norm.
As in the case of histogram models, assumption (**Ab'**) implies in the piecewise polynomial case assumption (**Ab**), see Section 3.5 of Chapter 3. Assumptions (**Aud**), ($\mathbf{Ad}_{\mathrm{Leb}}$) and (**Arpp**) allow us to guaranty the statements (**Alb**) and ($\mathbf{Ac}_\infty$) of the general set of assumptions in this case (see Lemmas 3.4 and 3.5 of Chapter 3). Moreover, we still have

$$R_{n,D_M} \propto \sqrt{\frac{D_M \ln n}{n}} \ ,$$

within a uniform constant over the models of $\mathcal{M}_n$. It is well-known that piecewise polynomials uniformly bounded in their degrees have good approximation properties in Besov spaces. More precisely, as stated in Lemma 12 of Barron, Birgé and Massart [13], if $\mathcal{X} = [0,1]$ and the regression function s_* belongs to the Besov space $B_{\alpha,p,\infty}(\mathcal{X})$ (see the definition in [13]), then taking models of piecewise polynomials of degree bounded by $r > \alpha - 1$ on regular partitions with respect to the Lebesgue measure Leb on $\mathcal{X}$, and assuming that P^X has a density with respect to Leb which is bounded in sup-norm, assumption ($\mathbf{Ap}_u$) is satisfied. It remains to find conditions in this context such that the lower bound on the bias in (**Ap**) is also satisfied.

Remark 4.1 *Since constants in the general set of assumptions (**GSA**) made above are uniform over the collection $\mathcal{M}_n$, we deduce from Theorem 3.1 applied with $\alpha = 2 + \alpha_{\mathcal{M}}$ and $A_- = A_+ = A_{\mathcal{M},+}$ that if assumptions (**P2**), (**Ab**), (**An**), (**Alb**) and ($\mathbf{Ac}_\infty$) hold, then a positive constant A_0 exists, depending on $\alpha_{\mathcal{M}}$, $A_{\mathcal{M},+}$ and on the constants A, $\sigma_{\min}$ and $r_{\mathcal{M}}$ defined in the general set of assumptions, such that for all $M \in \mathcal{M}_n$ satisfying*

$$0 < A_{\mathcal{M},+} (\ln n)^2 \leq D_M \ ,$$

by setting

$$\varepsilon_n(M) = A_0 \max\left\{ \left(\frac{\ln n}{D_M}\right)^{1/4} ; \ \left(\frac{D_M \ln n}{n}\right)^{1/4} ; \ \sqrt{R_{n,D_M}} \right\} \tag{4.15}$$

we have, for all $n \geq n_0(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}})$,

$$\mathbb{P}\left[(1 - \varepsilon_n(M)) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \leq P(Ks_n(M) - Ks_M) \leq (1 + \varepsilon_n(M)) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 10 n^{-2-\alpha_{\mathcal{M}}} \tag{4.16}$$

and

$$\mathbb{P}\left[\left(1 - \varepsilon_n^2(M)\right) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \leq P_n(Ks_M - Ks_n(M)) \leq \left(1 + \varepsilon_n^2(M)\right) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5 n^{-2-\alpha_{\mathcal{M}}} \ . \tag{4.17}$$

Moreover, for all $M \in \mathcal{M}_n$, we have by Theorem 3.2, for a positive constant A_u depending on $A, A_{cons}, r_{\mathcal{M}}$ and $\alpha_{\mathcal{M}}$ and for all $n \geq n_0(A_{cons}, n_1)$,

$$\mathbb{P}\left[P(Ks_n(M) - Ks_M) \geq A_u \frac{D_M \vee \ln n}{n} \right] \leq 3 n^{-2-\alpha_{\mathcal{M}}} \tag{4.18}$$

and

$$\mathbb{P}\left[P_n(Ks_M - Ks_n(M)) \geq A_u \frac{D_M \vee \ln n}{n} \right] \leq 3 n^{-2-\alpha_{\mathcal{M}}} \ . \tag{4.19}$$

The remainder of this chapter is devoted to the proofs.

4.4 Proofs

Before stating the proofs of Theorems 4.2 and 4.1, we need two technical lemmas. In the first lemma, we intend to evaluate the minimal penalty $\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$ for models of dimension not too large and not too small.

Lemma 4.1 *Assume (**P2**), (**Ab**), (**An**), (**Alb**) and (**Ac**$_\infty$) of the general set of assumptions defined in Section 4.3.1. Then, for every model $M \in \mathcal{M}_n$ of dimension D_M such that*

$$0 < A_{\mathcal{M},+}\left(\ln n\right)^2 \leq D_M \ ,$$

we have for all $n \geq n_0\left(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}}\right)$,

$$\left(1 - L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \tag{4.20}$$

$$\leq \left(1 + L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \ , \tag{4.21}$$

where $\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}; \left(\frac{D_M \ln n}{n}\right)^{1/4}; \sqrt{R_{n,D_M}}\right\}$ is defined in Remark 4.1.

Proof. As explained in Remark 4.1, under assumptions of Lemma 4.1 we can apply Theorem 3.1 with $A_- = A_+ = A_{\mathcal{M},+}$ and $\alpha = 2 + \alpha_{\mathcal{M}}$. For all $n \geq n_0\left(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}}\right)$, we thus have on an event $\Omega_1\left(M\right)$ of probability at least $1 - 5n^{-2-\alpha_{\mathcal{M}}}$,

$$\left(1 - \varepsilon_n\left(M\right)\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \leq P_n\left(Ks_M - Ks_n\left(M\right)\right) \leq \left(1 + \varepsilon_n\left(M\right)\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \ , \tag{4.22}$$

where $\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}; \ \left(\frac{D_M \ln n}{n}\right)^{1/4}; \ \sqrt{R_{n,D_M}}\right\}$. Moreover, as $|Y_i| \leq A$ *a.s.* and $\|s_M\|_\infty \leq A$ by (**Ab**), it holds

$$0 \leq P_n\left(Ks_M - Ks_n\left(M\right)\right) \leq P_n Ks_M = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - s_M\left(X_I\right)\right)^2 \leq 4A^2 \tag{4.23}$$

and as $D_M \geq 1$, we have

$$\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}; \ \left(\frac{D_M \ln n}{n}\right)^{1/4}; \ \sqrt{R_{n,D_M}}\right\} \geq A_0 n^{-1/8} \ . \tag{4.24}$$

We also have

$$\begin{aligned}&\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]\\ &= \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\mathbf{1}_{\Omega_1(M)}\right] + \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\mathbf{1}_{(\Omega_1(M))^c}\right] \ .\end{aligned} \tag{4.25}$$

Now notice that by (**An**) we have $\mathcal{K}_{1,M} \geq 2\sigma_{\min} > 0$. Hence, as $D_M \geq 1$, it comes from (4.23) and (4.24) that

$$0 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\mathbf{1}_{(\Omega_1(M))^c}\right] \leq 20A^2 n^{-2-\alpha_{\mathcal{M}}} \leq \frac{80A^2}{A_0^2\sigma_{\min}^2}\varepsilon_n^2\left(M\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \ . \tag{4.26}$$

Moreover, we have $\varepsilon_n(M) < 1$ for all $n \geq n_0(A_0, A_{\mathcal{M},+}, A_{cons})$, so by (4.22),

$$0 < \left(1 - 5n^{-2-\alpha_{\mathcal{M}}}\right)\left(1 - \varepsilon_n^2(M)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{\Omega_1(M)}\right] \tag{4.27}$$

$$\leq \left(1 - 5n^{-2-\alpha_{\mathcal{M}}}\right)\left(1 + \varepsilon_n^2(M)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \ . \tag{4.28}$$

Finally, noticing that $n^{-2-\alpha_{\mathcal{M}}} \leq A_0^{-2}\varepsilon_n^2(M)$ by (4.24), we use (4.26), (4.27) and (4.28) in (4.25) to conclude by straightforward computations that

$$L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}} = \frac{80A^2}{A_0^2\sigma_{\min}^2} + 5A_0^{-2} + 1$$

is convenient in (4.20) and (4.21), as A_0 only depends on $\alpha_{\mathcal{M}}$, $A_{\mathcal{M},+}$, A, $\sigma_{\min}$ and $r_{\mathcal{M}}$. ■

Lemma 4.2 *Let $\alpha > 0$. Assume that (**Ab**) of Section 4.3.1 is satisfied. Then a positive constant A_d exists, depending only in A, $A_{\mathcal{M},+}$, $\sigma_{\min}$ and α such that, by setting $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$, we have for all $M \in \mathcal{M}_n$,*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq A_d\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right)\right) \leq 2n^{-\alpha} \ . \tag{4.29}$$

*If moreover, assumptions (**P2**), (**Ab**), (**An**), (**Alb**) and (**Ac**$_\infty$) of the general set of assumptions defined in Section 4.3.1 hold, then for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^2 \leq D_M$ and for all $n \geq n_0(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha)$, we have*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + A_d\frac{\ln n}{\sqrt{D_M}}\mathbb{E}\left[p_2(M)\right]\right) \leq 2n^{-\alpha} \ , \tag{4.30}$$

where $p_2(M) := P_n\left(Ks_M - Ks_n(M)\right) \geq 0$.

Proof. We set

$$A_d = \max\left\{4A\sqrt{\alpha};\ \frac{8A^2}{3}\alpha;\ \frac{8A^2\alpha}{\sqrt{A_{\mathcal{M},+}\sigma_{\min}^2}} + \frac{16A^2\alpha}{3A_{\mathcal{M},+}\sigma_{\min}}\right\} \ . \tag{4.31}$$

Since by (**Ab**) we have $|Y| \leq A$ *a.s.* and $\|s_*\|_\infty \leq A$, it holds $\|s_*\|_\infty = \|\mathbb{E}[Y|X]\|_\infty \leq A$, and so $\|s_M - s_*\|_\infty \leq 2A$. Next, we apply Bernstein's inequality (7.46) to $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$. Notice that

$$K(s_M)(x,y) - K(s_*)(x,y) = \left(s_M(x) - s_*(x)\right)\left(s_M(x) + s_*(x) - 2y\right) \ ,$$

hence $\|Ks_M - Ks_*\|_\infty \leq 8A^2$. Moreover, as $\mathbb{E}[Y - s_*(X)|X] = 0$ and $\mathbb{E}\left[\left(Y - s_*(X)\right)^2|X\right] \leq \frac{(2A)^2}{4} = A^2$ we have

$$\begin{aligned}
&\mathbb{E}\left[\left(Ks_M(X,Y) - Ks_*(X,Y)\right)^2\right] \\
&= \mathbb{E}\left[\left(4\left(Y - s_*(X)\right)^2 + \left(s_M(X) - s_*(X)\right)^2\right)\left(s_M(X) - s_*(X)\right)^2\right] \\
&\leq 8A^2\mathbb{E}\left[\left(s_M(X) - s_*(X)\right)^2\right] \\
&= 8A^2\ell(s_*, s_M) \ ,
\end{aligned}$$

and therefore, by (7.46) we have for all $x > 0$,

$$\mathbb{P}\left(\left|\bar{\delta}\left(M\right)\right| \geq \sqrt{\frac{16A^2\ell\left(s_*, s_M\right)x}{n}} + \frac{8A^2x}{3n}\right) \leq 2\exp\left(-x\right) \ .$$

By taking $x = \alpha \ln n$, we then have

$$\mathbb{P}\left(\left|\bar{\delta}\left(M\right)\right| \geq \sqrt{\frac{16A^2\alpha\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{8A^2\alpha\ln n}{3n}\right) \leq 2n^{-\alpha} \ , \tag{4.32}$$

which gives the first part of Lemma 4.2 for A_d given in (4.31). Now, by noticing the fact that $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$, and by using it in (4.32) with $a = \ell\left(s_*, s_M\right)$, $b = \frac{4A^2\alpha\ln n}{n}$ and $\eta = D_M^{-1/2}$, we obtain

$$\mathbb{P}\left(\left|\bar{\delta}\left(M\right)\right| \geq \frac{\ell\left(s_*, s_M\right)}{\sqrt{D_M}} + \left(4\sqrt{D_M} + \frac{8}{3}\right)\frac{A^2\alpha\ln n}{n}\right) \leq 2n^{-\alpha} \ . \tag{4.33}$$

Then, for a model $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}\left(\ln n\right)^2 \leq D_M$, we apply Lemma 4.1 and by (4.20), it holds for all $n \geq n_0\left(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}}\right)$,

$$\left(1 - L_{A_{\mathcal{M},-},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[p_2\left(M\right)\right] \tag{4.34}$$

where $\varepsilon_n = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}, \ \left(\frac{D_M \ln n}{n}\right)^{1/4}, \ \sqrt{R_{n,D_M,\alpha}}\right\}$. Moreover as $D_M \leq A_{\mathcal{M},+} n\left(\ln n\right)^{-2}$ by (**P2**), $R_{n,D_M} \leq A_{cons}\left(\ln n\right)^{-1/2}$ by (4.6) and $A_{\mathcal{M},+}\left(\ln n\right)^2 \leq D_M$, we deduce that for all $n \geq n_0\left(A_{\mathcal{M},+}, A, A_{cons}, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}}\right)$,

$$L_{A_{\mathcal{M},-},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2\left(M\right) \leq 1/2 \ .$$

Now, since $\mathcal{K}_{1,M} \geq 2\sigma_{\min} > 0$ by (**An**), we have by (4.34), $\mathbb{E}\left[p_2\left(M\right)\right] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n}$ for all $n \geq n_0\left(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}}\right)$. This allows, using (4.33), to conclude the proof for the value of A_d given in (4.31) by simple computations. ■

In order to avoid cumbersome notations in the proofs of Theorems 4.2 and 4.1, when generic constants L and n_0 depend on constants defined in the general set of assumptions stated in Section 4.3.1, we will note $L_{(\mathbf{GSA})}$ and $n_0\left((\mathbf{GSA})\right)$.

Proof of Theorem 4.2. From the definition of the selected model $\widehat{M}$ given in (4.3), $\widehat{M}$ minimizes

$$\mathrm{crit}\left(M\right) := P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}\left(M\right) \ , \tag{4.35}$$

over the models $M \in \mathcal{M}_n$. Hence, $\widehat{M}$ also minimizes

$$\mathrm{crit}'\left(M\right) := \mathrm{crit}\left(M\right) - P_n\left(Ks_*\right) \ . \tag{4.36}$$

over the collection $\mathcal{M}_n$. Let us write

$$\begin{aligned}\ell\left(s_*, s_n\left(M\right)\right) &= P\left(Ks_n\left(M\right) - Ks_*\right)\\ &= P_n\left(Ks_n\left(M\right)\right) + P_n\left(Ks_M - Ks_n\left(M\right)\right) + \left(P_n - P\right)\left(Ks_* - Ks_M\right)\\ &\quad + P\left(Ks_n\left(M\right) - Ks_M\right) - P_n\left(Ks_*\right) \ .\end{aligned}$$

By setting

$$p_1(M) = P(Ks_n(M) - Ks_M) \ ,$$
$$p_2(M) = P_n(Ks_M - Ks_n(M)) \ ,$$
$$\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$$

and

$$\mathrm{pen}'_{\mathrm{id}}(M) = p_1(M) + p_2(M) - \bar{\delta}(M) \ ,$$

we have

$$\ell(s_*, s_n(M)) = P_n(Ks_n(M)) + p_1(M) + p_2(M) - \bar{\delta}(M) - P_n(Ks_*) \tag{4.37}$$

and by (4.36),

$$\mathrm{crit}'(M) = \ell(s_*, s_n(M)) + \left(\mathrm{pen}(M) - \mathrm{pen}'_{\mathrm{id}}(M)\right) \ . \tag{4.38}$$

As $\widehat{M}$ minimizes crit′ over $\mathcal{M}_n$, it is therefore sufficient by (4.38), to control $\mathrm{pen}(M) - \mathrm{pen}'_{\mathrm{id}}(M)$ - or equivalently $\mathrm{crit}'(M)$ - in terms of the excess risk $\ell(s_*, s_n(M))$, for every $M \in \mathcal{M}_n$, in order to derive oracle inequalities. Let Ω_n be the event on which:

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $A_{\mathcal{M},+}(\ln n)^3 \leq D_M$, (4.10) hold and

$$|p_1(M) - \mathbb{E}[p_2(M)]| \leq L_{(\mathbf{GSA})}\varepsilon_n(M)\,\mathbb{E}[p_2(M)] \tag{4.39}$$
$$|p_2(M) - \mathbb{E}[p_2(M)]| \leq L_{(\mathbf{GSA})}\varepsilon_n^2(M)\,\mathbb{E}[p_2(M)] \tag{4.40}$$
$$\left|\bar{\delta}(M)\right| \leq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + L_{(\mathbf{GSA})}\frac{\ln n}{\sqrt{D_M}}\mathbb{E}[p_2(M)] \tag{4.41}$$
$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{4.42}$$

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $D_M \leq A_{\mathcal{M},+}(\ln n)^3$, (4.11) holds together with

$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{4.43}$$
$$p_2(M) \leq L_{(\mathbf{GSA})}\frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA})}\frac{(\ln n)^3}{n} \tag{4.44}$$
$$p_1(M) \leq L_{(\mathbf{GSA})}\frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA})}\frac{(\ln n)^3}{n} \tag{4.45}$$

By (4.16), (4.17), (4.18) and (4.19) in Remark 4.1, Lemma 4.1, Lemma 4.2 applied with $\alpha = 2 + \alpha_{\mathcal{M}}$, and since (4.10) holds with probability at least $1 - A_p n^{-2}$, we get for all $n \geq n_0$ ((**GSA**)),

$$\mathbb{P}(\Omega_n) \geq 1 - A_p n^{-2} - 24\sum_{M \in \mathcal{M}_n} n^{-2-\alpha_{\mathcal{M}}} \geq 1 - L_{A_p, c_{\mathcal{M}}} n^{-2} \ .$$

Control on the criterion crit′ **for models of dimension not too small:**

We consider models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^3 \leq D_M$. Notice that (4.41) implies by (4.15) that, for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^3 \leq D_M$, for all $n \geq n_0$ ((**GSA**)),

$$\begin{aligned}\left|\bar{\delta}(M)\right| &\leq L_{(\mathbf{GSA})}\left(\frac{(\ln n)^3}{D_M}\cdot\frac{\ln n}{D_M}\right)^{1/4} \times \mathbb{E}[\ell(s_*, s_M) + p_2(M)] \\ &\leq L_{(\mathbf{GSA})}\varepsilon_n(M)\,\mathbb{E}[\ell(s_*, s_M) + p_2(M)] \ ,\end{aligned}$$

so that on Ω_n we have, for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$,

$$\begin{aligned}
&\left|\mathrm{pen}'_{\mathrm{id}} (M) - \mathrm{pen} (M)\right| \\
&\leq |p_1 (M) + p_2 (M) - \mathrm{pen} (M)| + \left|\bar{\delta} (M)\right| \\
&\leq |p_1 (M) + p_2 (M) - 2\mathbb{E} [p_2 (M)]| + \delta \mathbb{E} [p_2 (M)] + L_{(\mathbf{GSA})} \varepsilon_n (M) \mathbb{E} [\ell (s_*, s_M) + p_2 (M)] \\
&\leq L_{(\mathbf{GSA})} \varepsilon_n (M) \mathbb{E} [p_2 (M)] + \delta \mathbb{E} [p_2 (M)] + L_{(\mathbf{GSA})} \varepsilon_n (M) \mathbb{E} [\ell (s_*, s_M) + p_2 (M)] \\
&\leq \left(\delta + L_{(\mathbf{GSA})} \varepsilon_n (M)\right) \mathbb{E} [\ell (s_*, s_M) + p_2 (M)] \ . \qquad (4.46)
\end{aligned}$$

Now notice that using (**P2**) and (4.6) in (4.15) gives that for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0 ((\mathbf{GSA}))$, $0 < L_{(\mathbf{GSA})} \varepsilon_n (M) \leq \frac{1}{2}$. As $\ell (s_*, s_n (M)) = \ell (s_*, s_M) + p_1 (M)$, we thus have on Ω_n, for all $n \geq n_0 ((\mathbf{GSA}))$,

$$\begin{aligned}
0 &\leq \mathbb{E} [\ell (s_*, s_M) + p_2 (M)] \\
&\leq \ell (s_*, s_n (M)) + |p_1 (M) - \mathbb{E} [p_2 (M)]| \\
&\leq \ell (s_*, s_n (M)) + \frac{L_{(\mathbf{GSA})} \varepsilon_n (M)}{1 - L_{(\mathbf{GSA})} \varepsilon_n (M)} p_1 (M) \qquad \text{by (4.39)} \\
&\leq \frac{1 + L_{(\mathbf{GSA})} \varepsilon_n (M)}{1 - L_{(\mathbf{GSA})} \varepsilon_n (M)} \ell (s_*, s_n (M)) \\
&\leq \left(1 + L_{(\mathbf{GSA})} \varepsilon_n (M)\right) \ell (s_*, s_n (M)) \ . \qquad (4.47)
\end{aligned}$$

Hence, using (4.47) in (4.46), we have on Ω_n for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0 ((\mathbf{GSA}))$,

$$\left|\mathrm{pen}'_{\mathrm{id}} (M) - \mathrm{pen} (M)\right| \leq \left(\delta + L_{(\mathbf{GSA})} \varepsilon_n (M)\right) \ell (s_*, s_n (M)) \ . \qquad (4.48)$$

By consequence, for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0 ((\mathbf{GSA}))$, it holds on Ω_n, using (4.38) and (4.48),

$$\left(1 - \delta - L_{(\mathbf{GSA})} \varepsilon_n (M)\right) \ell (s_*, s_n (M)) \leq \mathrm{crit}' (M) \leq \left(1 + \delta + L_{(\mathbf{GSA})} \varepsilon_n (M)\right) \ell (s_*, s_n (M)) \ . \qquad (4.49)$$

Control on the criterion crit′ **for models of small dimension:**

We consider models $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$. By (4.11), (4.43) and (4.44), it holds on Ω_n, for any $\tau > 0$ and for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

$$\begin{aligned}
&\left|\mathrm{pen}'_{\mathrm{id}} (M) - \mathrm{pen} (M)\right| \\
&\leq p_1 (M) + p_2 (M) + \mathrm{pen} (M) + \left|\bar{\delta} (M)\right| \\
&\leq L_{(\mathbf{GSA})} \frac{(\ln n)^3}{n} + A_r \frac{(\ln n)^3}{n} + L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell (s_*, s_M) \ln n}{n}} + \frac{\ln n}{n}\right) \\
&\leq L_{(\mathbf{GSA}), A_r} \frac{(\ln n)^3}{n} + \tau \ell (s_*, s_M) + \left(\tau^{-1} + 1\right) L_{(\mathbf{GSA})} \frac{\ln n}{n} \\
&\leq L_{(\mathbf{GSA}), A_r} \frac{(\ln n)^3}{n} + \tau \ell (s_*, s_n (M)) + \left(\tau^{-1} + 1\right) L_{(\mathbf{GSA})} \frac{\ln n}{n} \ . \qquad (4.50)
\end{aligned}$$

Hence, by taking $\tau = (\ln n)^{-2}$ in (4.50) we get that for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$, it holds on Ω_n,

$$\left|\mathrm{pen}'_{\mathrm{id}} (M) - \mathrm{pen} (M)\right| \leq \frac{\ell (s_*, s_n (M))}{(\ln n)^2} + L_{(\mathbf{GSA}), A_r} \frac{(\ln n)^3}{n} \ . \qquad (4.51)$$

Moreover, by (4.38) and (4.51), we have on the event Ω_n, for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

$$\left(1 - (\ln n)^{-2}\right) \ell (s_*, s_n (M)) - L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \leq \mathrm{crit}' (M) \tag{4.52}$$

$$\leq \left(1 + (\ln n)^{-2}\right) \ell (s_*, s_n (M)) + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{4.53}$$

Oracle inequalities:

Recall that by the definition given in (4.2), an oracle model satisfies

$$M_* \in \arg \min_{M \in \mathcal{M}_n} \{\ell (s_*, s_n (M))\} \ . \tag{4.54}$$

By Lemmas 4.3 and 4.4 below, we control on Ω_n the dimensions of the selected model $\widehat{M}$ and the oracle model M_*. More precisely, by (4.66) and (4.68), we have on Ω_n, for any $\frac{1}{2} > \eta > (1 - \beta_+)_+ /2$ and for all $n \geq n_0 ((\mathbf{GSA}), \eta, \delta)$,

$$D_{\widehat{M}} \leq n^{1/2+\eta} \ , \tag{4.55}$$

$$D_{M_*} \leq n^{1/2+\eta} \ . \tag{4.56}$$

Now, from (4.55) we distinguish two cases in order to control $\mathrm{crit}' \left(\widehat{M}\right)$. If $A_{\mathcal{M},+} (\ln n)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta}$, we get by (4.49), for all $n \geq n_0 ((\mathbf{GSA}))$,

$$\mathrm{crit}' \left(\widehat{M}\right) \geq \left(1 - \delta - L_{(\mathbf{GSA})} \varepsilon_n \left(\widehat{M}\right)\right) \ell \left(s_*, s_n \left(\widehat{M}\right)\right) \ . \tag{4.57}$$

Otherwise, if $D_{\widehat{M}} \leq A_{\mathcal{M},+} (\ln n)^3$, we get by (4.52),

$$\left(1 - (\ln n)^{-2}\right) \ell \left(s_*, s_n \left(\widehat{M}\right)\right) - L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \leq \mathrm{crit}' \left(\widehat{M}\right) \ . \tag{4.58}$$

In all cases, we have by (4.57) and (4.58), for all $n \geq n_0 ((\mathbf{GSA}))$,

$$\mathrm{crit}' \left(\widehat{M}\right) \geq \left(1 - \delta - (\ln n)^{-2} - L_{(\mathbf{GSA})} \sup_{M \in \mathcal{M}_n, \ A_{\mathcal{M},+} (\ln n)^3 \leq D_M \leq n^{1/2+\eta}} \varepsilon_n (M)\right) \ell \left(s_*, s_n \left(\widehat{M}\right)\right)$$

$$- L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{4.59}$$

Similarly, from (4.56) we distinguish two cases in order to control $\mathrm{crit}' (M_*)$. If $A_{\mathcal{M},+} (\ln n)^3 \leq D_{M_*} \leq n^{1/2+\eta}$, we get by (4.49), for all $n \geq n_0 ((\mathbf{GSA}))$,

$$\mathrm{crit}' (M_*) \leq \left(1 + \delta + L_{(\mathbf{GSA})} \varepsilon_n (M_*)\right) \ell (s_*, s_n (M_*)) \ . \tag{4.60}$$

Otherwise, if $D_{M_*} \leq A_{\mathcal{M},+} (\ln n)^3$, we get by (4.53),

$$\mathrm{crit}' (M_*) \leq \left(1 + (\ln n)^{-2}\right) \ell (s_*, s_n (M_*)) + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{4.61}$$

In all cases, we deduce from (4.60) and (4.61) that we have for all $n \geq n_0 ((\mathbf{GSA}),\delta)$,

$$\mathrm{crit}' (M_*) \leq \left(1 + \delta + (\ln n)^{-2} + L_{(\mathbf{GSA})} \sup_{M \in \mathcal{M}_n, \ A_{\mathcal{M},+} (\ln n)^3 \leq D_M \leq n^{1/2+\eta}} \varepsilon_n (M)\right) \ell (s_*, s_n (M_*))$$

$$+ L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{4.62}$$

Hence, by setting

$$\theta_n = L_{(\mathbf{GSA})} \times \sup_{M \in \mathcal{M}_n,\ A_{\mathcal{M},+}(\ln n)^3 \le D_M \le n^{1/2+\eta}} \varepsilon_n(M) \ ,$$

we have by (4.15) and (4.6), for all $n \ge n_0((\mathbf{GSA}), \eta, \delta)$,

$$\theta_n \le \frac{L_{(\mathbf{GSA})}}{(\ln n)^{1/4}} \ , \quad (\ln n)^{-2} + \theta_n + \delta < 1 \ , \quad (\ln n)^{-2} + \theta_n < \frac{1-\delta}{2}$$

and we deduce from (4.59) and (4.62), since $\frac{1}{1-x} \le 1 + 2x$ for all $x \in \left[0, \frac{1}{2}\right)$, that for all $n \ge n_0((\mathbf{GSA}), \eta, \delta)$, it holds on Ω_n,

$$\begin{aligned} \ell\left(s_*, s_n\left(\widehat{M}\right)\right) &\le \left(\frac{1+\delta+(\ln n)^{-2}+\theta_n}{1-\delta-(\ln n)^{-2}-\theta_n}\right) \ell(s_*, s_n(M_*)) + \frac{L_{(\mathbf{GSA}),A_r}}{1-\delta-(\ln n)^{-2}-\theta_n} \frac{(\ln n)^3}{n} \\ &\le \left(\frac{1+\delta}{1-\delta} + \frac{5\left((\ln n)^{-2}+\theta_n\right)}{(1-\delta)^2}\right) \ell(s_*, s_n(M_*)) + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \end{aligned} \tag{4.63}$$

Inequality (4.13) is now proved.
It remains to prove the second part of Theorem 4.2. We assume that assumption (**Ap**) holds. From Lemmas 4.3 and 4.4, we have that for any $\frac{1}{2} > \eta > (1-\beta_+)_+/2$ and for all $n \ge n_0((\mathbf{GSA}), C_-, \beta_-, \eta, \delta)$, it holds on Ω_n,

$$A_{\mathcal{M},+}(\ln n)^3 \le D_{\widehat{M}} \le n^{1/2+\eta} \ , \tag{4.64}$$

$$A_{\mathcal{M},+}(\ln n)^3 \le D_{M_*} \le n^{1/2+\eta} \ . \tag{4.65}$$

Now, using (4.57) and (4.60), by the same kind of computations leading to (4.63), we deduce that it holds on Ω_n, for all $n \ge n_0((\mathbf{GSA}), C_-, \beta_-, \eta, \delta)$,

$$\begin{aligned} \ell\left(s_*, s_n\left(\widehat{M}\right)\right) &\le \left(\frac{1+\delta+\theta_n}{1-\delta-\theta_n}\right) \ell(s_*, s_n(M_*)) \\ &\le \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{(1-\delta)^2}\right) \ell(s_*, s_n(M_*)) \ . \end{aligned}$$

Thus inequality (4.14) is proved and Theorem 4.2 follows. ■

Lemma 4.3 (Control on the dimension of the selected model) *Assume that the general set of assumptions (**GSA**) hold. Let $\eta > (1-\beta_+)_+/2$. If $n \ge n_0((\mathbf{GSA}), \eta, \delta)$ then, on the event Ω_n defined in the proof of Theorem 4.2, it holds*

$$D_{\widehat{M}} \le n^{1/2+\eta} \ . \tag{4.66}$$

*If moreover (**Ap**) holds, then for all $n \ge n_0((\mathbf{GSA}), C_-, \beta_-, \eta, \delta)$, we have on the event Ω_n,*

$$A_{\mathcal{M},+}(\ln n)^3 \le D_{\widehat{M}} \le n^{1/2+\eta} \ . \tag{4.67}$$

Lemma 4.4 (Control on the dimension of oracle models) *Assume that the general set of assumptions (**GSA**) hold. Let $\eta > (1-\beta_+)_+/2$. If $n \ge n_0((\mathbf{GSA}), \eta)$ then, on the event Ω_n defined in the proof of Theorem 4.2, it holds*

$$D_{M_*} \le n^{1/2+\eta} \ . \tag{4.68}$$

*If moreover (**Ap**) holds, then for all $n \ge n_0((\mathbf{GSA}), C_-, \beta_-, \eta)$, we have on the event Ω_n,*

$$A_{\mathcal{M},+}(\ln n)^3 \le D_{M_*} \le n^{1/2+\eta} \ . \tag{4.69}$$

Proof of Lemma 4.3. Recall that $\widehat{M}$ minimizes

$$\mathrm{crit}'(M) = \mathrm{crit}(M) - P_n K s_* = \ell(s_*, s_M) - p_2(M) + \bar{\delta}(M) + \mathrm{pen}(M) \tag{4.70}$$

over the models $M \in \mathcal{M}_n$.

1. Lower bound on $\mathrm{crit}'(M)$ for small models in the case where (**Ap**) hold : let $M \in \mathcal{M}_n$ be such that $D_M < A_{\mathcal{M},+}(\ln n)^3$. We then have on Ω_n,

$$\begin{aligned}
\ell(s_*, s_M) &\geq C_- A_{\mathcal{M},+}^{-\beta_-} (\ln n)^{-3\beta_-} \quad \text{by (\textbf{Ap})} \\
\mathrm{pen}(M) &\geq 0 \\
p_2(M) &\leq L_{(\mathbf{GSA})} \frac{(\ln n)^3}{n} \qquad \text{from (4.44)} \\
\bar{\delta}(M) &\geq -L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right) \text{ from (4.43).}
\end{aligned}$$

Since by (**Ab**), we have $0 \leq \ell(s_*, s_M) \leq 4A^2$, we deduce that for all $n \geq n_0((\mathbf{GSA}), C_-, \beta_-)$,

$$\mathrm{crit}'(M) \geq \frac{C_- A_{\mathcal{M},+}^{-\beta_-}}{2} (\ln n)^{-3\beta_-} \ . \tag{4.71}$$

2. Lower bound for large models : let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/2+\eta}$. From (4.10) and (4.40) we have on Ω_n,

$$\mathrm{pen}(M) - p_2(M) \geq \left(1 - \delta - L_{(\mathbf{GSA})} \varepsilon_n^2(M)\right) \mathbb{E}\left[p_2(M)\right] \ .$$

Using (**P2**), (4.6) and the fact that $D_M \geq n^{1/2+\eta}$ in (4.15), we deduce that for all $n \geq n_0((\mathbf{GSA}), \eta, \delta)$, $L_{(\mathbf{GSA})} \varepsilon_n^2(M) \leq \frac{1}{2}(1-\delta)$ and as by (**An**), $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ we also deduce from Lemma 4.1 that for all $n \geq n_0((\mathbf{GSA}), \eta)$, $\mathbb{E}\left[p_2(M)\right] \geq \frac{\sigma_{\min}^2}{2} \frac{D_M}{n}$. By consequence, it holds for all $n \geq n_0((\mathbf{GSA}), \eta, \delta)$,

$$\mathrm{pen}(M) - p_2(M) \geq \frac{\sigma_{\min}^2}{4} (1-\delta) \frac{D_M}{n} \ . \tag{4.72}$$

From (4.42) it holds on Ω_n,

$$\bar{\delta}(M) \geq -L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right) \ . \tag{4.73}$$

Hence, as $D_M \geq n^{1/2+\eta}$ and as by (**Ab**), $0 \leq \ell(s_*, s_M) \leq 4A^2$, we deduce from (4.70), (4.72) and (4.73) that we have on Ω_n, for all $n \geq n_0((\mathbf{GSA}), \eta, \delta)$,

$$\mathrm{crit}'(M) \geq (1-\delta) L_{(\mathbf{GSA})} n^{-1/2+\eta} \ . \tag{4.74}$$

3. A better model exists for $\mathrm{crit}'(M)$: from (**P3**), there exists $M_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n}$. Then, for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$A_{\mathcal{M},+}(\ln n)^3 \leq \sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n} \leq n^{1/2+\eta} \ .$$

Using ($\mathbf{Ap}_u$),

$$\ell(s_*, s_{M_0}) \leq C_+ n^{-\beta_+/2} \ . \tag{4.75}$$

By (4.41), we have on Ω_n, for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$\left|\bar{\delta}(M_0)\right| \leq \frac{\ell(s_*, s_{M_0})}{\sqrt{D_{M_0}}} + L_{(\mathbf{GSA})}\frac{\ln n}{\sqrt{D_{M_0}}}\mathbb{E}[p_2(M_0)] \tag{4.76}$$

and by (4.10),

$$\mathrm{pen}(M_0) \leq 3\mathbb{E}[p_2(M_0)] \ .$$

Hence, as $\mathcal{K}_{1,M} \leq 6A$ and $\ell(s_*, s_{M_0}) \leq 4A^2$ by (**Ab**) and as for all $n \geq n_0((\mathbf{GSA}))$ $\varepsilon_n(M) \leq 1$, we deduce from inequalities (4.75), (4.76) and Lemma 4.1 that for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$\left|\bar{\delta}(M_0)\right| \leq L_{(\mathbf{GSA})}\left(n^{-(\beta_+/2+1/4)} + \ln(n)\, n^{-3/4}\right)$$

and

$$\mathrm{pen}(M_0) \leq L_{(\mathbf{GSA})} n^{-1/2} \ .$$

By consequence, we have on Ω_n, for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$\begin{aligned}\mathrm{crit}'(M_0) &\leq \ell(s_*, s_{M_0}) + \left|\bar{\delta}(M_0)\right| + \mathrm{pen}(M_0) \\ &\leq L_{(\mathbf{GSA})}\left(n^{-\beta_+/2} + n^{-1/2}\right) \ .\end{aligned} \tag{4.77}$$

To conclude, notice that the upper bound (4.77) is smaller than the lower bound given in (4.74) for all $n \geq n_0((\mathbf{GSA}), \eta, \delta)$. Hence, points 2 and 3 above yield inequality (4.66). Moreover, the upper bound (4.77) is smaller than lower bounds given in (4.71), derived by using (**Ap**), and (4.74), for all $n \geq n_0((\mathbf{GSA}), C_-, \beta_-, \eta, \delta)$. This thus gives (4.67) and Lemma 4.3 is proved. ■

Proof of Lemma 4.4. By definition, M_* minimizes

$$\ell(s_*, s_n(M)) = \ell(s_*, s_M) + p_1(M)$$

over the models $M \in \mathcal{M}_n$.

1. Lower bound on $\ell(s_*, s_n(M))$ for small models : let $M \in \mathcal{M}_n$ be such that $D_M < A_{\mathcal{M},+}(\ln n)^3$. In this case we have

$$\ell(s_*, s_n(M)) \geq \ell(s_*, s_M) \geq C_- A_{\mathcal{M},+}^{-\beta_-}(\ln n)^{-3\beta_-} \text{ by } (\mathbf{Ap}). \tag{4.78}$$

2. Lower bound of $\ell(s_*, s_n(M))$ for large models : let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/2+\eta}$. From (4.39) we get on Ω_n,

$$p_1(M) \geq \left(1 - L_{(\mathbf{GSA})}\varepsilon_n(M)\right)\mathbb{E}[p_2(M)] \ .$$

Using (**P2**), (4.6) and the fact that $D_M \geq n^{1/2+\eta}$ in (4.15), we deduce that for all $n \geq n_0((\mathbf{GSA}), \eta)$, $L_{(\mathbf{GSA})}\varepsilon_n(M) \leq \frac{1}{2}$ and as by (**An**), $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ we also deduce from Lemma 4.1 that for all $n \geq n_0((\mathbf{GSA}), \eta)$, $\mathbb{E}[p_2(M)] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n}$. By consequence, it holds for all $n \geq n_0((\mathbf{GSA}), \eta)$, on the event Ω_n,

$$\ell(s_*, s_n(M)) \geq p_1(M) \geq \frac{\sigma_{\min}^2}{4}\frac{D_M}{n} \geq \frac{\sigma_{\min}^2}{4} n^{-1/2+\eta} \ . \tag{4.79}$$

3. A better model exists for $\ell(s_*, s_n(M))$: from (**P3**), there exists $M_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n}$. Moreover, for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$A_{\mathcal{M},+}(\ln n)^3 \leq \sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n} \leq n^{1/2+\eta} \ .$$

Using $(\mathbf{Ap}_u)$,

$$\ell(s_*, s_{M_0}) \leq C_+ n^{-\beta_+/2}$$

and by (4.39)

$$p_1(M_0) \leq \left(1 + L_{(\mathbf{GSA})}\varepsilon_n(M)\right) \mathbb{E}[p_2(M_0)]$$

Hence, as $\mathcal{K}_{1,M} \leq 6A$ by $(\mathbf{Ab})$ and as, by (4.6) and (4.15), for all $n \geq n_0((\mathbf{GSA}))$ it holds $\varepsilon_n(M) \leq 1$, we deduce from Lemma 4.1 that for all $n \geq n_0((\mathbf{GSA}))$, on the event Ω_n,

$$p_1(M_0) \leq L_{(\mathbf{GSA})}\frac{D_M}{n} \leq L_{(\mathbf{GSA})}n^{-1/2} .$$

By consequence, on Ω_n, for all $n \geq n_0((\mathbf{GSA}))$,

$$\begin{aligned}\ell(s_*, s_n(M_0)) &= \ell(s_*, s_{M_0}) + p_1(M_0) \\ &\leq L_{(\mathbf{GSA})}\left(n^{-\beta_+/2} + n^{-1/2}\right) . \end{aligned} \tag{4.80}$$

The upper bound (4.80) is smaller than the lower bound (4.79) for all $n \geq n_0((\mathbf{GSA}), \eta)$, and this gives (4.68). If $(\mathbf{Ap})$ hold, then the upper bound (4.80) is smaller than the lower bounds (4.78) and (4.79) for all $n \geq n_0((\mathbf{GSA}), C_-, \beta_-, \eta)$, which proves (4.69) and allows to conclude the proof of Lemma 4.4. ■

Proof of Theorem 4.1. Similarly to the proof of Theorem 4.2, we consider the event Ω'_n of probability at least $1 - L_{c_{\mathcal{M}}, A_p}n^{-2}$ for all $n \geq n_0((\mathbf{GSA}))$, on which: (4.8) holds and

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $A_{\mathcal{M},+}(\ln n)^2 \leq D_M$ it holds

$$|p_1(M) - \mathbb{E}[p_2(M)]| \leq L_{(\mathbf{GSA})}\varepsilon_n(M)\mathbb{E}[p_2(M)] , \tag{4.81}$$

$$|p_2(M) - \mathbb{E}[p_2(M)]| \leq L_{(\mathbf{GSA})}\varepsilon_n^2(M)\mathbb{E}[p_2(M)] . \tag{4.82}$$

- For all models $M \in \mathcal{M}_n$ with $D_M \leq A_{\mathcal{M},+}(\ln n)^2$ it holds

$$p_2(M) \leq L_{(\mathbf{GSA})}\frac{(\ln n)^2}{n} . \tag{4.83}$$

- For every $M \in \mathcal{M}_n$,

$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) . \tag{4.84}$$

Let $d \in (0,1)$ to be chosen later.

Lower bound on $D_{\widehat{M}}$. Remind that $\widehat{M}$ minimizes

$$\mathrm{crit}'(M) = \mathrm{crit}(M) - P_n K s_* = \ell(s_*, s_M) - p_2(M) + \bar{\delta}(M) + \mathrm{pen}(M) . \tag{4.85}$$

1. Lower bound on $\mathrm{crit}'(M)$ for "small" models : assume that $M \in \mathcal{M}_n$ and

$$D_M \leq dA_{rich}n(\ln n)^{-2} .$$

We have

$$\ell(s_*, s_M) + \mathrm{pen}(M) \geq 0 \tag{4.86}$$

and from (4.84), as $\ell(s_*, s_M) \leq 4A^2$ by (**Ab**), we get on Ω'_n, for all $n \geq n_0((\mathbf{GSA}),d)$,

$$\begin{aligned}\bar{\delta}(M) &\geq -L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right)\\ &\geq -L_{(\mathbf{GSA})}\sqrt{\frac{\ln n}{n}}\\ &\geq -d \times A^2 A_{rich}(\ln n)^{-2} \ . \end{aligned} \tag{4.87}$$

Then, if $D_M \geq A_{\mathcal{M},+}(\ln n)^2$, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab**) and as, by (4.6) and (4.15), for all $n \geq n_0((\mathbf{GSA}))$ it holds $L_{(\mathbf{GSA})}\varepsilon_n(M) \leq 1$, we deduce from (4.82) and Lemma 4.1 that for all $n \geq n_0((\mathbf{GSA}))$,

$$p_2(M) \leq 2\mathbb{E}[p_2(M)] \leq 36A^2\frac{D_M}{n} \leq d \times 36A^2 A_{rich}(\ln n)^{-2} \ .$$

Whenever $D_M \leq A_{\mathcal{M},+}(\ln n)^2$, (4.83) gives that, for all $n \geq n_0((\mathbf{GSA}),d)$, on the event Ω'_n,

$$p_2(M) \leq L_{(\mathbf{GSA})}\frac{(\ln n)^2}{n} \leq d \times 36A^2 A_{rich}(\ln n)^{-2} \ .$$

Hence, we have checked that for all $n \geq n_0((\mathbf{GSA}),d)$, on the event Ω'_n,

$$-p_2(M) \geq -d \times 36A^2 A_{rich}(\ln n)^{-2} \ , \tag{4.88}$$

and finally, by using (4.86), (4.87) and (4.88) in (4.85), we deduce that on Ω'_n, for all $n \geq n_0((\mathbf{GSA}),d)$,

$$\mathrm{crit}'(M) \geq -d \times 37A^2 A_{rich}(\ln n)^{-2} \ . \tag{4.89}$$

2. There exists a better model for $\mathrm{crit}'(M)$: By (**P3**), for all $n \geq n_0(A_{\mathcal{M},+}, A_{rich})$ a model $M_1 \in \mathcal{M}_n$ exists such that

$$A_{\mathcal{M},+}(\ln n)^2 \leq \frac{A_{rich}n}{(\ln n)^2} \leq D_{M_1} \ .$$

We then have on Ω'_n,

$$\begin{aligned}\ell(s_*, s_{M_1}) &\leq A_{rich}^{-\beta_+}(\ln n)^{2\beta_+} n^{-\beta_+} && \text{by } (\mathbf{Ap}_u)\\ p_2(M_1) &\geq \left(1 - L_{(\mathbf{GSA})}\varepsilon_n^2(M_1)\right)\mathbb{E}[p_2(M_1)] && \text{by (4.82)}\\ \mathrm{pen}(M_1) &\leq A_{\mathrm{pen}}\mathbb{E}[p_2(M_1)] && \text{by (4.8)}\\ \left|\bar{\delta}(M_1)\right| &\leq L_{(\mathbf{GSA})}\sqrt{\frac{\ln n}{n}} && \text{by (4.84) and } (\mathbf{Ab})\end{aligned}$$

and therefore,

$$\mathrm{crit}'(M_1) \leq \left(-1 + A_{\mathrm{pen}} + L_{(\mathbf{GSA})}\varepsilon_n^2(M_1)\right)\mathbb{E}[p_2(M_1)] + L_{(\mathbf{GSA})}\sqrt{\frac{\ln n}{n}} + A_{rich}^{-\beta_+}\frac{(\ln n)^{2\beta_+}}{n^{\beta_+}} \ . \tag{4.90}$$

Hence, as $-1 + A_{\mathrm{pen}} < 0$, and as by (4.6), (4.15), (**An**) and Lemma 4.1 it holds for all $n \geq n_0((\mathbf{GSA}), A_{\mathrm{pen}})$

$$L_{(\mathbf{GSA})}\varepsilon_n^2(M_1) \leq \frac{1 - A_{\mathrm{pen}}}{2} \quad \text{and} \quad \mathbb{E}[p_2(M_1)] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n} \geq \frac{\sigma_{\min}^2 A_{rich}}{2}(\ln n)^{-2} \ ,$$

we deduce from (4.90) that on Ω'_n, for all $n \geq n_0((\mathbf{GSA}), A_{\mathrm{pen}})$,

$$\mathrm{crit}'(M_1) \leq -\frac{1}{4}(1 - A_{\mathrm{pen}})\sigma_{\min}^2 A_{rich}(\ln n)^{-2} \ . \tag{4.91}$$

Now, by taking

$$0 < d = \left(\frac{1}{149} \left(1 - A_{\text{pen}}\right) \left(\frac{\sigma_{\min}}{A} \right)^2 \right) \wedge \frac{1}{2} < 1 \tag{4.92}$$

and by comparing (4.89) and (4.91), we deduce that on Ω'_n, for all $n \geq n_0\left((\mathbf{GSA}), A_{\text{pen}}\right)$, for all $M \in \mathcal{M}_n$ such that $D_M \leq dA_{rich} n \left(\ln n\right)^{-2}$,

$$\text{crit}'\left(M_1\right) < \text{crit}'\left(M\right)$$

and so

$$D_{\widehat{M}} > dA_{rich} n \left(\ln n\right)^{-2} \ . \tag{4.93}$$

Excess Risk of $s_n\left(\widehat{M}\right)$. We take d with the value given in (4.92). First notice that for all $n \geq n_0\left(A_{\mathcal{M},+}, A_{rich}, d\right)$, we have $dA_{rich} n \left(\ln n\right)^{-2} \geq A_{\mathcal{M},+} \left(\ln n\right)^2$. Hence, for all $M \in \mathcal{M}_n$ such that $D_M \geq dA_{rich} n \left(\ln n\right)^{-2}$, by (4.6), (4.15), (**P2**), (**An**) and Lemma 4.1, it holds on Ω'_n for all $n \geq n_0\left((\mathbf{GSA}), A_{\text{pen}}\right)$, using (4.81),

$$\ell\left(s_*, s_n\left(M\right)\right) \geq p_1\left(M\right) \geq \frac{\sigma_{\min}^2}{2} \frac{D_M}{n} \geq \frac{d\sigma_{\min}^2 A_{rich}}{2} \left(\ln n\right)^{-2} \ .$$

By (4.93), we thus get that on Ω'_n, for all $n \geq n_0\left((\mathbf{GSA}), A_{\text{pen}}\right)$,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \frac{d\sigma_{\min}^2 A_{rich}}{2} \left(\ln n\right)^{-2} \ . \tag{4.94}$$

Moreover, the model M_0 defined in (**P3**) satisfies, for all $n \geq n_0\left((\mathbf{GSA})\right)$,

$$A_{\mathcal{M},+} \left(\ln n\right)^3 \leq \sqrt{n} \leq D_{M_0} \leq c_{rich} \sqrt{n}$$

and so using $(\mathbf{Ap}_u)$,

$$\ell\left(s_*, s_{M_0}\right) \leq C_+ n^{-\beta_+/2} \ .$$

In addition, by (4.39),

$$p_1\left(M\right) \leq \left(1 + L_{(\mathbf{GSA})} \varepsilon_n\left(M\right)\right) \mathbb{E}\left[p_2\left(M\right)\right] \ .$$

Hence, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab**) and as, by (4.6) and (4.15), for all $n \geq n_0\left((\mathbf{GSA})\right)$ it holds $\varepsilon_n\left(M\right) \leq 1$, we deduce from Lemma 4.1 that for all $n \geq n_0\left((\mathbf{GSA})\right)$

$$p_1\left(M\right) \leq L_{(\mathbf{GSA})} \frac{D_M}{n} \leq L_{(\mathbf{GSA})} n^{-1/2} \ .$$

By consequence, for all $n \geq n_0\left((\mathbf{GSA})\right)$,

$$\ell\left(s_*, s_n\left(M_0\right)\right) \leq L_{(\mathbf{GSA})} \left(n^{-\beta_+/2} + n^{-1/2} \right) \tag{4.95}$$

and the ratio between the two bounds (4.94) and (4.95) is larger than $\ln\left(n\right)$ for all $n \geq n_0\left(L_{(\mathbf{GSA})}, A_{\text{pen}}\right)$, which yields (4.9). ■

Chapitre 5

Slope Heuristics and nonasymptotic optimality of AIC criterion for penalized maximum likelihood on histograms

5.1 Introduction

This chapter is devoted to the study of some penalized maximum likelihood model selection procedures for the estimation of density on histograms. There is a huge amount of literature on the problem of model selection by penalized maximum likelihood criteria, even in the more restrictive question of selecting an histogram, that goes back to Akaike's pioneer work. In the early seventies, Akaike [2] proposed to select a model by penalizing the empirical likelihood of maximum likelihood estimators by the number of parameters in each model. The analysis of Akaike [2] on the model selection procedure defined by the so-called Akaike's Information Criterion (AIC), is fundamentally asymptotic in the sense that the author considers a given finite collection of models with the number of data going to infinity. This asymptotic setting is irrelevant in many situations and thus many efforts have been made to develop nonasymptotic analysis of model selection procedures, letting the dimension of the models and the cardinality of the collection of models depend on the number of data. As pointed out by Boucheron and Massart [27], it is nevertheless worth mentioning that early works of Akaike [3] and Mallows [59] in model selection relied, although in a disguised form, on the Wilks' phenomenon (Wilks [88]) that asserts that in smooth parametric density estimation the difference between the maximum likelihood and the likelihood of the sampling distribution converges towards a chi-square distribution where the number of degrees of freedom coincides with the model dimension. This phenomenon has been generalized by Boucheron and Massart [27] in a nonasymptotic way, considering the empirical excess risk in a M-estimation with bounded contrast setting, and is actually one the main results supporting the conjecture that the slope heuristics introduced by Birgé and Massart [23] hold in some general framework, see Arlot and Massart [10]. Let us now describe some works related to the selection of maximum likelihood estimators.

Barron and Sheu [15] give some risks bounds on maximum likelihood estimation considering sequences of regular exponential families made of polynomials, splines and trigonometric series. They achieve an accurate trade-off between the bias term and the variance term considering that log-density functions have square integrable derivatives. Considering general models, Barron, Birgé and Massart [13] give strategies of penalization in a nonasymptotic framework and derive oracle inequalities for the Hellinger risk. In particular, the considered penalty terms take into account the complexity of the collection of models, but as a prize to pay for generality, they

involve absolute constants that may be unrealistic.

Particularizing the structure of the models to histograms, Castellan [30] proposes a modified Akaike's criterion that also takes into account the complexity of the collection of models, and that lead to significant changes compared to AIC criterion in the case of large collections of irregular partitions. She derives nonasymptotic oracle inequalities for the Hellinger and Kullback-Leibler risks of the selected model, with leading constants in front of the oracle only depending on the multiplicative constant in the penalty term and being optimized for a penalty term corresponding to AIC in the case of regular histograms. But, despite the fact that she gives optimal controls from above and from below for the mean of the Hellinger and Kullback-Leibler risks on a fixed model (see Proposition 2.4 and 2.6 in [30]), the derived oracle inequalities are not sufficiently sharp to recover the asymptotic optimality of AIC in the case of regular histograms, as the leading constants are bounded away from one even if the number of data is going to infinity. Castellan [30] also give a lower bound for the penalty term that corresponds to half AIC penalty, when the unknown density is uniform on the unit interval and the partitions are regular. This result seems to indicate that the slope heuristics exhibited by Birgé and Massart [23] is satisfied in the context of maximum likelihood estimation of density, at least when the considered models are regular histograms. Castellan [31] has also been able to generalize her study to exponential models where the logarithm of functions are piecewise polynomials. By distinguishing between regular and irregular partitions defining the models, she gives significant bounds in Hellinger risk for procedures of model selection based on a modified Akaike's criterion. We also refer to the introduction of Castellan [30] for a state of the art on the problem of selecting histograms, and in particular the related question of optimal cell width in the case of regular histograms.

We show in this chapter that the slope heuristics is valid when the collection of models is of polynomial complexity with respect to the number of data and the considered partitions satisfy some lower regularity assumption. More precisely, we identify the minimal penalty as half AIC penalty. For a penalty function less than the minimal one, we show that the procedure of model selection totally misbehaves in the sense that the Kullback-Leibler excess risk of the selected model is much larger than the oracle one, and the selected dimension is systematically large too. On the contrary, when the penalty function is larger than the minimal one, assuming that the bias of the models are bounded from above and from below by a power of the number of elements in each partition, we show a nonasymptotic pathwise oracle inequality for the Kullback-Leibler excess risk of the selected model. The assumption on the bias of the models holds true when the unknown density is a non constant α-Hölder function. Moreover, if the penalty function is close to two times the minimal one, the leading constant in the oracle inequality is close to one, and is even converging to one when the number of data is going to infinity, meaning that we are close to the optimal penalty. This allows us to show nonasymptotic quasi-optimality of AIC in this context. From a practical point of view, as our results theoretically validate the data-driven calibration of penalty exposed by Arlot and Massart in [10] and as the penalty shape is known in this case and is equal to the dimension of the models, we are able to provide a data-driven model selection procedure that asymptotically behaves like AIC procedure. Moreover, this data-driven procedure should perform better than AIC for small numbers of data. A simulation study about this fact is still in progress.

Our analysis, that significantly differs from Castellan's approach in [30], is based on the concept of regular contrast exposed in Chapter 7, which is shown to be satisfied on each histogram model. Indeed, on each model, the Kullback-Leibler divergence with respect to the Kullback-Leibler projection of the unknown density is shown to be close to a weighted $L_2(P)$ norm, locally around the Kullback-Leibler projection, where P is the sampling distribution. Our approach then relies on two central facts : under a lower regularity assumption on the partitions, the models are equipped with a localized basis structure, and assuming moreover that the unknown density is uniformly bounded from above, the maximum likelihood estimators

are consistent in sup-norm, uniformly over the collection of models, and converge towards their corresponding Kullback-Leibler projections. We notice that this notion of convergence in sup-norm, which is essential in our methodology, is also present in the work of Castellan, slightly disguised in the term $\Omega_m(\varepsilon)$ defined in Section 2.3 of [30].

Finally, histogram models of densities combine two properties : on the one hand they are a particular case of exponential models, and on the other hand they can be viewed as the subset of positive functions in an affine space. Our approach is based on the second property, whereas Castellan's one relies on the first property, taking advantage of the linear structure of the contrasted functions. We conjecture that the slope phenomenon discovered by Birgé and Massart in a generalized linear Gaussian model setting can be extended in the two directions described above. In each case, one of the main task will be to prove the consistency in sup-norm of the maximum likelihood estimators on the considered models, as further explained in Section 5.4.

The chapter is organized as follows. In Section 5.2 we describe the statistical framework, the considered models and we investigate in Section 5.2.3 the regular structure of the Kullback-Leibler contrast on histogram models. We state in Section 5.3 our main results. In Section 5.4 we give arguments concerning possible developments of the two possible generalizations described above. The proofs are postponed to the end of the chapter.

5.2 Framework and notations

5.2.1 Maximum Likelihood Estimation

We assume that we have n i.i.d. observations $(\xi_1, ..., \xi_n)$ with common unknown law P on a measurable space $(\mathcal{Z}, \mathcal{T})$ and that ξ is a generic random variable of law P on $(\mathcal{Z}, \mathcal{T})$ and independent of the sample $(\xi_1, ..., \xi_n)$. We also assume that there exists a known probability measure μ on $(\mathcal{Z}, \mathcal{T})$ such that P admits a density s_* with respect to μ :

$$s_* = \frac{dP}{d\mu} .$$

Our goal is to estimate the density s_*.
For a measurable suitable integrable function f on $\mathcal{Z}$, we set

$$Pf = P(f) = \mathbb{E}[f(\xi)]$$
$$\mu f = \mu(f) = \int_{\mathcal{Z}} f d\mu$$

and if

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$$

denote the empirical distribution associated to the data $(\xi_1, ..., \xi_n)$,

$$P_n f = P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(\xi_i) .$$

Moreover, taking the convention $\ln(0) = -\infty$ and defining the positive part as $(x)_+ = x \vee 0$, we set

$$\mathcal{S} = \left\{ s : \mathcal{Z} \longrightarrow \mathbb{R}_+ \; ; \; \int_{\mathcal{Z}} s d\mu = 1 \text{ and } P(\ln s)_+ < +\infty \right\} . \tag{5.1}$$

We assume in the sequel that the unknown density s_* belongs to $\mathcal{S}$. In fact, in order to derive our results, we will assume in Section 5.3 that s_* is uniformly bounded away from zero and

uniformly upper bounded on $\mathcal{Z}$. For now, note that since $P(\ln s_*)_- < +\infty$ and $s_* \in \mathcal{S}$, we have $P|\ln(s_*)| < +\infty$. Moreover, the Kullback-Leibler contrast K is defined on $\mathcal{S}$ to be

$$K : s \in \mathcal{S} \longmapsto (z \in \mathcal{Z} \longmapsto -\ln(s(z)))$$

and thus the risk

$$PK(s) = P(Ks) = PKs = P(\ln s)_- - P(\ln s)_+$$

as well as the excess risk

$$\ell(s_*, s) = P(Ks) - P(Ks_*) = P(Ks - Ks_*)$$

are well defined on $\mathcal{S}$ and can be possibly infinite. Now, for two probability distributions P_s and P_t on $(\mathcal{Z}, \mathcal{T})$ of respective densities s and t with respect to μ, the Kullback-Leibler divergence of P_t with respect to P_s is defined to be

$$\mathcal{K}(P_s, P_t) = \begin{cases} \int_{\mathcal{Z}} \ln\left(\frac{dP_s}{dP_t}\right) dP_t = \int_{\mathcal{Z}} s \ln\left(\frac{s}{t}\right) d\mu & \text{if } P_s \ll P_t \\ +\infty & \text{otherwise.} \end{cases} \tag{5.2}$$

By misuse of notation we will denote $\mathcal{K}(s,t)$ rather than $\mathcal{K}(P_s, P_t)$ and by Jensen inequality we notice that $\mathcal{K}(s,t)$ is a nonnegative quantity, equal to zero if and only if $s = t$ μ-$a.s.$ Hence, for any $s \in \mathcal{S}$, the excess risk $\ell(s_*, s)$ satisfies

$$\begin{aligned} \ell(s_*, s) &= P(Ks - Ks_*) \\ &= \int_{\mathcal{Z}} \ln\left(\frac{s_*}{s}\right) s_* d\mu \\ &= \mathcal{K}(s_*, s) \geq 0 \end{aligned} \tag{5.3}$$

and this nonnegative quantity is equal to zero if and only if $s_* = s$ $\mu - a.s.$ We thus deduce that the unknown density s_* is uniquely defined by

$$\begin{aligned} s_* &= \arg\min_{s \in \mathcal{S}} \{P(-\ln s)\} \\ &= \arg\min_{s \in \mathcal{S}} \{PK(s)\} \ . \end{aligned} \tag{5.4}$$

For a subset $\widetilde{M} \subset \mathcal{S}$, we define the maximum likelihood estimator on $\widetilde{M}$, whenever it exists, by

$$\begin{aligned} s_n\left(\widetilde{M}\right) &\in \arg\min_{s \in \widetilde{M}} \{P_n K s\} \\ &= \arg\min_{s \in \widetilde{M}} \left\{ \frac{1}{n} \sum_{i=1}^{n} -\ln(s(\xi_i)) \right\} \ . \end{aligned} \tag{5.5}$$

Finally, for any $s \in L_2(P)$, we denote by

$$\|s\|_2 = \left(\int_{\mathcal{Z}} s^2 dP \right)^{1/2}$$

its quadratic norm.

5.2.2 Histogram models

The models $\widetilde{M}$ that we consider here to define the maximum likelihood estimators as in (5.5) are subsets of linear spaces M made of histograms. More precisely, for a finite partition Λ_M of cardinality $|\Lambda_M| = D_M$, we set

$$M = \left\{ s = \sum_{I \in \Lambda_M} \beta_I \mathbf{1}_I \; ; \; \beta = (\beta_I)_{I \in \Lambda_M} \in \mathbb{R}^{D_M} \right\}$$

the linear vector space of piecewise constant functions with respect to Λ_M and we assume that any element I of the partition Λ_M is of positive measure with respect to μ :

$$\text{for all } I \in \Lambda_M, \quad \mu(I) > 0 \; . \tag{5.6}$$

By misuse of language, the space M is also called "model" or "histogram model". The linear dimension of M is equal to D_M. In addition we associate to the model M the subset $\widetilde{M}$ of the functions in M that are densities with respect to μ,

$$\widetilde{M} = \left\{ s \in M \; ; \; s \geq 0 \text{ and } \int_{\mathcal{Z}} s d\mu = 1 \right\} \; .$$

As the partition Λ_M is finite, we have $P(\ln s)_+ < +\infty$ for all $s \in \widetilde{M}$ and so $\widetilde{M} \subset \mathcal{S}$. Hence, by (5.5), we can associate to $\widetilde{M}$ the maximum likelihood estimator $s_n\left(\widetilde{M}\right)$ and in the following we denote it $s_n(M)$ rather than $s_n\left(\widetilde{M}\right)$. We state in the next proposition some well-known properties that are satisfied by histogram models submitted to the procedure of maximum likelihood estimation (see for example Massart [61], Section 7.3).

Proposition 5.1 *Let*

$$s_M = \sum_{I \in \Lambda_M} \frac{P(I)}{\mu(I)} \mathbf{1}_I \; . \tag{5.7}$$

Then $s_M \in \widetilde{M}$ and s_M is called the Kullback-Leibler projection of s_* onto $\widetilde{M}$. Moreover, it holds

$$s_M = \arg\min_{s \in \widetilde{M}} P(Ks) \; . \tag{5.8}$$

The following Pythagorean-like identity for the Kullback-Leibler divergence holds, for every $s \in \widetilde{M}$,

$$\mathcal{K}(s_*, s) = \mathcal{K}(s_*, s_M) + \mathcal{K}(s_M, s) \; . \tag{5.9}$$

We also have the following formula

$$s_n(M) = \sum_{I \in \Lambda_M} \frac{P_n(I)}{\mu(I)} \mathbf{1}_I \; , \tag{5.10}$$

and so the maximum likelihood estimator on M is well defined and corresponds to the classical histogram estimator of s_* associated to the partition Λ_M.

Remark 5.1 *Histogram models are special cases of general exponential families exposed for example in Barron and Sheu [15] (see also Castellan [31] for the case of exponential models of piecewise polynomials). The projection property (5.9) can be generalized to exponential models (see Lemma 3 of [15] and Csiszár [34]).*

Remark 5.2 *As by (5.3) we have*

$$P\left(Ks_M - Ks_*\right) = \mathcal{K}\left(s_*, s_M\right)$$

and for any $s \in \widetilde{M}$,

$$P\left(Ks - Ks_*\right) = \mathcal{K}\left(s_*, s\right)$$

we easily deduce from (5.9) that the excess risk on $\widetilde{M}$ *is still a Kullback-Leibler divergence, as we then have for any* $s \in \widetilde{M}$,

$$P\left(Ks - Ks_M\right) = \mathcal{K}\left(s_M, s\right) \ . \tag{5.11}$$

Moreover it is easy to see using (5.10) that the maximum likelihood estimator on a histogram model M *is also the least-squares estimator.*

As explained in Chapter 7, we shall ask for a particular analytical structure of the considered models in order to derive sharp upper and lower bounds for the excess risk on each model of reasonable dimension. Namely, we require here that the models are fulfilled with a localized basis structure with respect to the $L_2\left(P\right)$ norm. As stated in the following lemma, this property is available when the unknown density of data is uniformly bounded away from zero and when the partition Λ_M related to the model M satisfies some lower regularity property with respect to the measure of reference μ.

Lemma 5.1 *Let* $A_{\min}, A_\Lambda > 0$. *Let* Λ_M *be some finite partition of* $\mathcal{Z}$ *and* M *be the model of piecewise constant functions on the partition* Λ_M. *Assume that*

$$\inf_{z \in \mathcal{Z}} s_*\left(z\right) \geq A_{\min} > 0 \quad \text{and} \quad D_M \inf_{I \in \Lambda_M} \mu\left(I\right) \geq A_\Lambda > 0 \ . \tag{5.12}$$

Set $r_M = \left(A_{\min} A_\Lambda\right)^{-1/2}$ *and define, for all* $I \in \Lambda_M$,

$$\varphi_I = \left(P\left(I\right)\right)^{-1/2} \mathbf{1}_I \ .$$

Then the family $\left(\varphi_I\right)_{I \in \Lambda_M}$ *is an orthonormal basis of* $\left(M, L_2\left(P\right)\right)$ *that satisfies, for all* $\beta = \left(\beta_I\right)_{I \in \Lambda_M} \in \mathbb{R}^{D_M}$,

$$\left\| \sum_{I \in \Lambda_M} \beta_I \varphi_I \right\|_\infty \leq r_M \sqrt{D_M}\, |\beta|_\infty \tag{5.13}$$

where $|\beta|_\infty = \max\left\{|\beta_I| \ , \ I \in \Lambda_M\right\}$. *As a consequence,*

$$\sup_{s \in M, \ \|s\|_2 \leq 1} \|s\|_\infty \leq r_M \sqrt{D_M} \ . \tag{5.14}$$

The proof of Lemma 5.1 is straightforward and can be found in Section 5.5.1.

5.2.3 Regularity of the Kullback-Leibler contrast

Our goal is to study the performance of maximum likelihood estimators, that we measure by their excess risk. So we are interested in the random quantity $P\left(Ks_n\left(M\right) - Ks_*\right)$. Moreover, since we can write

$$P\left(Ks_n\left(M\right) - Ks_*\right) = P\left(Ks_n\left(M\right) - Ks_M\right) + P\left(Ks_M - Ks_*\right)$$

and since the bias $P\left(Ks_M - Ks_*\right)$ is deterministic, we focus on the quantity

$$P\left(Ks_n\left(M\right) - Ks_M\right) \geq 0 \ ,$$

that we want to bound in probability. We will often call this last quantity the excess risk of the estimator on M or the true excess risk of $s_n(M)$, by opposition to the empirical excess risk for which the expectation is taken over the empirical measure : $P_n(Ks_M - Ks_n(M)) \geq 0$.
We notice that by Proposition 5.1, the excess risk of the maximum likelihood estimator on M is still a Kullback-Leibler divergence if M is a model of histograms, as we have

$$P(Ks_n(M) - Ks_M) = \mathcal{K}(s_M, s_n(M)) \ .$$

The following lemma provides an expansion of the contrast around s_M on M as the sum of a linear part and a second order part which behaves as a quadratic. This is an example of what we call more generally a regular contrast, see Section 2.2 of Chapter 2.

Lemma 5.2 *Assume that*

$$\inf_{z\in\mathcal{Z}} s_*(z) \geq A_{\min} > 0 \tag{5.15}$$

and consider $s \in \widetilde{M}$ *such that*

$$\|s - s_M\|_\infty < A_{\min} \ . \tag{5.16}$$

Then we have $\inf_{z\in\mathcal{Z}} s(z) > 0$ *and it holds for all* $z \in \mathcal{Z}$*,*

$$(Ks)(z) - (Ks_M)(z) = \psi_{1,M}(z)(s - s_M)(z) + \psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right) \tag{5.17}$$

with

$$\psi_{1,M}(z) = -\frac{1}{s_M(z)}$$

and, for all $t \in (-1, +\infty)$*,*

$$\psi_2(t) = t - \ln(1+t) \ .$$

The two following lemmas ensure that the remainder term $\psi_2\left(\left(\frac{s-s_M}{s_M}\right)(z)\right)$ in the expansion of the contrast (5.17) indeed behaves like a quadratic term, when the unknown density is uniformly bounded from below and elements $s - s_M$ are sufficiently small in sup-norm.

Lemma 5.3 *Let* $\delta \in [0, A_{\min}/2]$*. Assume that*

$$\inf_{z\in\mathcal{Z}} s_*(z) \geq A_{\min} > 0 \ . \tag{5.18}$$

Then, for all $z \in \mathcal{Z}$ *and* $s \in \widetilde{M}$ *such that* $|(s - s_M)(z)| \leq \delta$ *, it holds*

$$\left|\left(\frac{s - s_M}{s_M}\right)(z)\right| \leq \frac{\delta}{A_{\min}} \leq \frac{1}{2}$$

and for all $(x, y) \in \left[-\frac{\delta}{A_{\min}}, \frac{\delta}{A_{\min}}\right]$*,*

$$|\psi_2(x) - \psi_2(y)| \leq \frac{2\delta}{A_{\min}}|x - y| \ . \tag{5.19}$$

Lemma 5.3 allows us in the Technical Lemmas of Section 5.5.5 to apply the contraction principle given in Theorem 7.4 of Chapter 7 in order to control the second order terms.
Now, the following lemma states that if s is close to s_M in sup-norm, then the Kullback-Leibler divergence is close to a weighted $L_2(P)$ norm.

Lemma 5.4 *Assume that*

$$\inf_{z\in\mathcal{Z}} s_*(z) \geq A_{\min} > 0 \ . \tag{5.20}$$

Let $\delta > 0$ *such that*

$$0 < \delta \leq \frac{A_{\min}}{2} \ .$$

Then for all $s \in M$ *such that* $\|s - s_M\|_\infty \leq \delta$, *we have* $\inf_{z\in\mathcal{Z}} s(z) > 0$, *and if moreover* $\int_{\mathcal{Z}} s d\mu = 1$ *then* $s \in \widetilde{M}$ *and it holds*

$$\left(\frac{1}{2} - \frac{2\delta}{3A_{\min}}\right) \left\| \frac{s - s_M}{s_M} \right\|_2^2 \leq \mathcal{K}(s_M, s) = P(Ks - Ks_M) \leq \left(\frac{1}{2} + \frac{2\delta}{3A_{\min}}\right) \left\| \frac{s - s_M}{s_M} \right\|_2^2 \ . \tag{5.21}$$

The proofs of Lemmas 5.3 and 5.4 are postponed to Section 5.5.1.

5.3 Results

We state here our main results. In Section 5.3.1, we investigate the convergence in sup-norm of the histogram estimators towards the Kullback-Leibler projections. This will be needed to derive the sharp upper and lower bounds in probability for the true and empirical excess risks of Section 5.3.2. Finally, the results obtained in a model selection framework are stated in Section 5.3.3.

5.3.1 Rates of convergence in sup-norm of histogram estimators

In order to handle second order terms in the expansion of the contrast (5.17) we show that the histogram estimator $s_n(M)$ is consistent in sup-norm towards the Kullback-Leibler projection s_M. More precisely, for models having a not too large dimension, the following lemma ensures the convergence in sup-norm of $s_n(M)$ towards s_M at the rate

$$R_{n,D_M} \propto \sqrt{\frac{D_M \ln n}{n}} \ .$$

Proposition 5.2 *Let* α, $A_+, A_*, A_\Lambda > 0$. *Consider the linear model* M *of histograms defined on a finite partition* Λ_M *of* $\mathcal{Z}$, *with* $|\Lambda_M| = D_M$ *its linear dimension. Assume*

$$\|s_*\|_\infty \leq A_* < +\infty \ , \tag{5.22}$$

$$D_M \inf_{I\in\Lambda_M} \mu(I) \geq A_\Lambda > 0 \ , \tag{5.23}$$

and

$$D_M \leq A_+ \frac{n}{(\ln n)^2} \leq n \ .$$

Then a positive constant A_c *exists, only depending on* A_Λ, A_*, A_+ *and* α *such that*

$$\mathbb{P}\left[\|s_n(M) - s_M\|_\infty \geq A_c \sqrt{\frac{D_M \ln n}{n}} \right] \leq 2n^{-\alpha} \ . \tag{5.24}$$

In Proposition 5.2, we need to assume that the target s_* is uniformly bounded from above over $\mathcal{Z}$, in order to derive the consistency in sup-norm of the histogram estimator towards the Kullback-Leibler projection s_M. This rather strong assumption can be avoided by normalizing the difference between the histogram estimator and the Kullback-Leibler projection by the

latter quantity. The rate of convergence of the sup-norm of the normalized difference is the same as in Proposition 5.2, that is

$$\sqrt{\frac{D_M \ln n}{n}} \ ,$$

but we assume in Proposition 5.3 that the target s_* is uniformly bounded away from zero over $\mathcal{Z}$.

Proposition 5.3 *Let α, $A_+, A_{\min}, A_\Lambda > 0$. Consider the linear model M of histograms defined on a finite partition Λ_M of $\mathcal{Z}$, with $|\Lambda_M| = D_M$ its linear dimension. Assume*

$$\inf_{z \in \mathcal{Z}} s_* (z) \geq A_{\min} > +\infty \ , \tag{5.25}$$

$$D_M \inf_{I \in \Lambda_M} \mu (I) \geq A_\Lambda > 0 \ , \tag{5.26}$$

and

$$D_M \leq A_+ \frac{n}{(\ln n)^2} \leq n \ .$$

Then a positive constant A_c exists, only depending on A_Λ, $A_{\min}, A_+$ and α such that

$$\mathbb{P}\left[\left\| \frac{s_n (M) - s_M}{s_M} \right\|_\infty \geq A_c \sqrt{\frac{D_M \ln n}{n}} \right] \leq 2 n^{-\alpha} \ . \tag{5.27}$$

As claimed in Remark 5.3 below, Proposition 5.3 indeed suffices in the proof of Theorem 5.1 to handle the second order terms appearing in the expansion of the contrast (5.17).
The proof of Proposition 5.2 can be found in Section 5.5.2.

5.3.2 True and empirical risks bounds

In this section, we fix the linear model M made of histograms and we are interested by upper and lower bounds for the true excess risk $P (K s_n (M) - K s_M)$ on M and for its empirical counterpart $P_n (K s_M - K s_n (M))$. We show that under reasonable assumptions the true excess risk is equivalent to the empirical one, which is one of the keystones to prove the slope phenomenon and the optimality of AIC that we state in Section 5.3.3.

Theorem 5.1 *Let $\alpha, A_+, A_-, A_{\min}, A_*, A_\Lambda > 0$ and let M be a linear model of histograms defined on a finite partition Λ_M. The finite dimension of M is denoted by D_M. Assume that*

$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_* (z) \ , \tag{5.28}$$

$$\| s_* \|_\infty \leq A_* < +\infty \ , \tag{5.29}$$

$$0 < A_\Lambda \leq D_M \inf_{I \in \Lambda_M} \mu (I) \tag{5.30}$$

and

$$0 < A_- (\ln n)^2 \leq D_M \leq A_+ \frac{n}{(\ln n)^2} \leq n \ .$$

Then a positive constant A_0 exists, only depending on $\alpha, A_-, A_+, A_*, A_{\min}$ and A_Λ, such that by setting

$$\varepsilon_n (M) = A_0 \max \left\{ \left(\frac{\ln n}{D_M} \right)^{1/4}, \left(\frac{D_M \ln n}{n} \right)^{1/4} \right\} \ , \tag{5.31}$$

we have, for all $n \geq n_0\left(A_+, A_-, A_{\min}, A_*, \alpha\right)$,

$$\mathbb{P}\left[P\left(Ks_n\left(M\right) - Ks_M\right) \geq \left(1 - \varepsilon_n\left(M\right)\right)\frac{D_M - 1}{2n}\right] \geq 1 - 6n^{-\alpha}\ , \tag{5.32}$$

$$\mathbb{P}\left[P\left(Ks_n\left(M\right) - Ks_M\right) \leq \left(1 + \varepsilon_n\left(M\right)\right)\frac{D_M - 1}{2n}\right] \geq 1 - 6n^{-\alpha}\ , \tag{5.33}$$

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right) \geq \left(1 - \varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n}\right] \geq 1 - 2n^{-\alpha}\ , \tag{5.34}$$

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right) \leq \left(1 + \varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n}\right] \geq 1 - 4n^{-\alpha}\ . \tag{5.35}$$

In the previous Theorem we achieve sharp upper and lower bounds for the true and empirical excess risk on M. They are optimal at the first order since the leading constants are equal in upper and lower bounds. Moreover, Theorem 5.1 establishes the equivalence with high probability of the true and empirical excess risks for models of reasonable dimension.
Castellan [30] also asks for a lower regularity property of the partition, for example in Proposition 2.5 where she derive a sharp control of the Kullback-Leibler excess risk of the histogram estimator on a fixed model. More precisely she assumes that there exists a positive constant B such that

$$\inf_{I \in \Lambda_M} \mu\left(I\right) \geq B\frac{\left(\ln n\right)^2}{n}\ . \tag{5.36}$$

This latter assumption is thus weaker than (5.30) for the considered model as its dimension D_M is less than the order $n\left(\ln n\right)^{-2}$. We could assume (5.36) instead of (5.30) in order to derive Theorem 5.1. This would lead to less precise results for second order terms in the deviations of the excess risks but the first order bounds would be preserved. More precisely, if we replace assumption (5.30) in Theorem 5.1 by Castellan's assumption (5.36), a careful look at the proofs of Lemma 5.1, Proposition 5.2 and Theorem 5.1 show that the conclusions of Theorem 5.1 are still valid for

$$\varepsilon_n = A_0\left(\ln n\right)^{-1/4}$$

where A_0 is some positive constant. Thus assumption (5.30) is not a fundamental restriction in comparison to Castellan's work [30], but it leads to more precise results in terms of deviations of the true and empirical excess risks of the histogram estimator.

Remark 5.3 *In the proof of Theorem 5.1 given in Section 5.5.3 and relying on the technical lemmas given in Section 5.5.5, we localize the analysis on the subset*

$$B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) = \left\{s \in M, \left\|s - s_M\right\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\}\ ,$$

where $\tilde{R}_{n,D_M,\alpha} = A_\infty\sqrt{D_M n^{-1}\ln n}$ *is defined in (5.78). This is possible by using Propostion 5.2, which states the convergence of* $\left\|s_n\left(M\right) - s_M\right\|_\infty$ *towards zero at a rate proportional to* $\sqrt{D_M n^{-1}\ln n}$ *with high probability. Considering Proposition 5.3, where we establish the convergence of* $\left\|\left(s_n\left(M\right) - s_M\right)/s_M\right\|_\infty$ *towards zero, again at a rate proportional to* $\sqrt{D_M n^{-1}\ln n}$ *with high probability, we can rather localize the analysis on the subset*

$$\left\{s \in M, \left\|\frac{s - s_M}{s_M}\right\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\}\ .$$

The gain is that in Proposition 5.3 - on contrary to Proposition 5.2 - we do not have to assume that the target s_* *is uniformly bounded from above over* $\mathcal{Z}$*. Hence, a careful look at the proof of Theorem 5.1, and especially at the proofs of Lemmas 5.2, 5.3 and 5.4 given in Section 5.5.1 and the proofs of Lemmas 5.9, 5.10, 5.11 and 5.12 given in Section 5.5.5, show that we can make*

straightforward modifications in order to recover results of Theorem 5.1 - with different values of the constants - without the assumption (5.29) of uniform boundedness of the target s_* on $\mathcal{Z}$. More precisely, the other assumptions of Theorem 5.1 would stay the same, and assumption (5.29) would be replaced by the much weaker moment condition

$$P\left(\ln s_*\right)_+ < +\infty \ ,$$

ensuring that $s_* \in \mathcal{S}$. The same remark apply to Theorem 5.2 below.

We turn now to upper bounds in probability for the true and empirical excess risks on models with small dimensions. Our aim here is not to compute sharp constants. In fact, information given by Theorem 5.2 suffices to our needs as we use it in the proofs of the results stated in Section 5.3.3 in order to control model selection procedures for small models.

Theorem 5.2 *Let $\alpha, A_+, A_{\min}, A_*, A_\Lambda > 0$ and let M be a linear model of histograms defined on a finite partition Λ_M. The finite dimension of M is denoted by D_M. Assume that*

$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_*\left(z\right) \ , \tag{5.37}$$

$$\left\|s_*\right\|_\infty \leq A_* < +\infty \ , \tag{5.38}$$

$$0 < A_\Lambda \leq D_M \inf_{I \in \Lambda_M} \mu\left(I\right) \tag{5.39}$$

and

$$1 \leq D_M \leq A_+ \frac{n}{\left(\ln n\right)^2} \leq n \ .$$

Then a positive constant A_u exists, only depending on $\alpha, A_+, A_*, A_{\min}, A_\Lambda$, such that for all $n \geq n_0\left(A_+, A_*, A_{\min}, A_\Lambda, \alpha\right)$,

$$\mathbb{P}\left[P\left(Ks_n\left(M\right) - Ks_M\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \ . \tag{5.40}$$

and

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \ . \tag{5.41}$$

The proofs of Theorems 5.1 and 5.2 can be found in Section 5.5.3.

5.3.3 Model Selection

We study in this section the behavior of model selection procedures by penalization of histogram estimators of the density s_*. Under reasonable assumptions stated below, we derive in Theorem 5.4 a pathwise oracle inequality for the Kullback-Leibler excess risk of the selected estimator, with constant almost one in front of the excess risk of the oracle when the penalty is close to Akaike's one. Our result thus establishes in this case the nonasymptotic quasi-optimality of AIC procedure with respect to the Kullback-Leibler risk. This is an improvement of results of Castellan [30] in the case of "small" collections of models.
Moreover, we validate the slope heuristics first formulated by Birgé and Massart [23] and extended by Arlot and Massart [10]. Indeed, we show in Theorem 5.3 that if the chosen penalty is less than half of Akaike's penalty then the model selection procedure totally misbehaves. More precisely, the excess risk of the selected estimator is much bigger than the one of the oracle and the dimension of the selected model also explode. This jump of dimension can be exploited in practice to derive a data-driven procedure of calibration of AIC penalty, as explained in Arlot and Massart [10]. This improvement should lead to better performances, at least when the number of data is "small". A comparison, based on simulations, of AIC

procedure and the calibration of the linear shape of the optimal penalty via the slope heuristics is still in progress.
Let us now define the model selection procedure. Given a collection of models $\mathcal{M}_n$ with cardinality depending on the number of data n and its associated collection of maximum likelihood estimators

$$\{s_n(M); M \in \mathcal{M}_n\} \ ,$$

and a nonnegative penalty function pen on $\mathcal{M}_n$

$$\text{pen} : M \in \mathcal{M}_n \longmapsto \text{pen}(M) \in \mathbb{R}^+$$

the output of the procedure, also called the selected model is

$$\widehat{M} \in \arg\min_{M \in \mathcal{M}_n} \{P_n(Ks_n(M)) + \text{pen}(M)\}. \tag{5.42}$$

The target of the model selection procedure is

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \{P(Ks_n(M))\}$$

and the associated M-estimator $s_n(M_*)$ is called an oracle. Let us now state the set of assumptions.

Set of assumptions (SA)

(P1) Polynomial complexity of $\mathcal{M}_n$: $\text{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(P2) Upper bound on dimensions of models in $\mathcal{M}_n$: there exists a positive constant $A_{\mathcal{M},+}$ such that for every $M \in \mathcal{M}_n$,

$$D_M \leq A_{\mathcal{M},+} \frac{n}{(\ln n)^2} \leq n \ . \tag{5.43}$$

(P3) Richness of $\mathcal{M}_n$: there exist $M_0, M_1 \in \mathcal{M}_n$ such that $D_{M_0} \in [\sqrt{n}, c_{rich}\sqrt{n}]$ and $D_{M_1} \geq A_{rich} n (\ln n)^{-2}$.

(Abd) The unknown density s_* is uniformly bounded from below and from above : there exist some positive finite constants $A_{\min}, A_*$ such that,

$$\|s_*\|_\infty \leq A_* < \infty \tag{5.44}$$

and

$$\inf_{z \in \mathcal{Z}} s_*(z) \geq A_{\min} > 0 \ . \tag{5.45}$$

($\mathbf{Ap}_u$) The bias decreases as a power of D_M : there exist $\beta_+ > 0$ and $C_+ > 0$ such that

$$\ell(s_*, s_M) \leq C_+ D_M^{-\beta_+} \ .$$

(Alr) Lower regularity of the partition with respect to μ : A positive finite constant A_Λ such that, for all $M \in \mathcal{M}_n$,

$$D_M \inf_{I \in \Lambda_M} \mu(I) \geq A_\Lambda > 0 \ . \tag{5.46}$$

Theorem 5.3 *Under the set of assumptions (**SA**) defined above, we further assume that for* $A_{\text{pen}} \in [0,1)$ *and* $A_p > 0$, *we have with probability at least* $1 - A_p n^{-2}$, *for all* $M \in \mathcal{M}_n$,

$$0 \leq \text{pen}(M) \leq A_{\text{pen}} \frac{D_M - 1}{2n} \ . \tag{5.47}$$

Then there exist two positive constants A_1, A_2 *independent of* n *such that, with probability at least* $1 - A_1 n^{-2}$, *we have for* $n \geq n_0((\boldsymbol{SA}), A_{\text{pen}})$,

$$D_{\widehat{M}} \geq A_2 n \ln(n)^{-2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \ln(n) \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n(M)\right)\right\} \ .$$

In Theorem 5.3 stated above we prove the existence of a minimal penalty, which is half of AIC. It thus validate the first part of the slope heuristics. Moreover, by Theorem 5.1 of Section 5.3.2, we see that for models of dimension not too small we have, with high probability,

$$P_n\left(Ks_M - Ks_n(M)\right) \approx \frac{D_M - 1}{2n} \ .$$

In fact, a careful look at the proof of Theorem 5.3 - which follows from arguments that are essentially the same as those of the proof of Theorem 4.1 - shows that, by Lemma 5.6 of Section 5.5.4, we can replace the condition (5.47) by the following one,

$$0 \leq \text{pen}(M) \leq A_{\text{pen}} \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] \ .$$

This latter formulation is also interesting because it presents our results as a particular case of the general statement of the slope heuristics given by Arlot and Massart in [10].

Theorem 5.4 *Assume that the set of assumptions (**SA**) hold together with*

(Ap) *The bias decreases like a power of* D_M *: there exist* $\beta_- \geq \beta_+ > 0$ *and* $C_+, C_- > 0$ *such that*

$$C_- D_M^{-\beta_-} \leq \ell(s_*, s_M) \leq C_+ D_M^{-\beta_+} \ .$$

Moreover, for $\delta \in \left(0, \frac{1}{2}\right)$ *and* $L > 0$, *assume that an event of probability at least* $1 - A_p n^{-2}$ *exists on which, for every model* $M \in \mathcal{M}_n$ *such that* $D_M \geq A_{\mathcal{M},+} (\ln n)^2$,

$$(1-\delta)\frac{D_M - 1}{n} \leq \text{pen}(M) \leq (1+\delta)\frac{D_M - 1}{n} \ . \tag{5.48}$$

Then, for $\frac{1}{2} > \eta > (1 - \beta_+)_+ / 2$, *there exists a constant* A_3 *and a sequence*

$$\theta_n = \sup_{M \in \mathcal{M}_n} \left\{\varepsilon_n(M) \ ; \ A_{\mathcal{M},+} (\ln n)^3 \leq D_M \leq n^{\eta + 1/2}\right\} \leq \frac{L_{(\boldsymbol{SA})}}{(\ln n)^{1/4}}$$

such that with probability at least $1 - A_3 n^{-2}$, *it holds for all* $n \geq n_0((\boldsymbol{SA}), C_-, \beta_-, \eta, \delta)$,

$$D_{\widehat{M}} \leq n^{\eta + 1/2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+2\delta}{1-2\delta} + \frac{5\theta_n}{(1-2\delta)^2}\right) \ell\left(s_*, s_n(M_*)\right) \ . \tag{5.49}$$

Theorem 5.4 states that if the penalty is more than half AIC for models of reasonable dimension then the model selection procedure achieve a nonasymptotic oracle inequality. Moreover, we prove the nonasymptotic quasi-optimality of the selected histogram estimator when the empirical excess risk is penalized by Akaike's criterion, which corresponds to the case where $\delta = 0$. Indeed, we derive in (5.49) a nonasymptotic pathwise oracle inequality with leading constant almost one. So Theorem 5.4 validates the second part of the slope heuristics. In order to recover the general formulation of the slope heuristics given by Arlot and Massart, we could replace the condition (5.48) by the following one

$$2(1-\delta)\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] \leq \mathrm{pen}(M) \leq 2(1+\delta)\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right]$$

and the conclusions of the theorem would be exactly the same.
The proofs of Theorems 5.3 and 5.4 can be found in Section 5.5.4.

Comments

Let us now comment on the set of assumptions (**SA**) . Assumption (**P1**) states that the collection of models has a "small" complexity, more precisely a polynomially increasing one. For this kind of complexities, if one wants to perform a good model selection procedure for prediction, the chosen penalty should estimate the mean of the ideal one on each model. Indeed, as Talagrand's type inequalities for the empirical process are pre-Gaussian, they allow to neglect the deviations of the quantities of interest from their mean, uniformly over the collection of models. This is not the case for too large collection of models, where one has to put an extra-log factor depending the complexity of the collection of models inside the penalty, see for example [30] and Massart [61].
The assumption (5.45) stating that the unknown density is uniformly bounded by below can also be found in the work of Castellan [30]. The author assumes moreover in Theorem 3.4 where she derives an oracle inequality for the Kullback-Leibler excess risk of the histogram estimator, that the target is of finite sup-norm as in inequality (5.44). But in the case of the Hellinger risk this assumption is replaced in [30] by the weaker assumption that the logarithm of the unknown density s_* is square integrable with respect to the sampling distribution.
In assumption (**P3**) we assume that we have a model M_0 of reasonable dimension and a model M_1 of high dimension. We demand in ($\mathbf{Ap}_u$) that the quality of approximation of the collection of models is good enough in terms of bias. More precisely, we require a polynomially decreasing of excess risk of Kullback-Leibler projections of the unknown density onto the models. For a density s_* uniformly bounded away from zero, this is satisfied when for example, $\mathcal{Z}$ is the unit interval, $\mu = \mathrm{Leb}$ is the Lebesgue measure on the unit interval, the partitions Λ_M are regular and the density s_* belongs to the set $\mathcal{H}(H,\alpha)$ of α-hölderian functions for some $\alpha \in (0,1]$: if $f \in \mathcal{H}(H,\alpha)$, then for all $(x,y) \in \mathcal{Z}^2$

$$|f(x) - f(y)| \leq H|x-y|^{\alpha} \ .$$

In that case, $\beta_+ = 2\alpha$ is convenient and when the chosen penalty is more than half AIC in our case, the procedure is adaptive to the parameters H and α, see Castellan [30].
In assumption (**Ap**) of Theorem 5.4 we also assume that the bias $\ell(s_*, s_M)$ is lower bounded by a power of the dimension D_M of the model M. This hypothesis is in fact quite classical as it has been used by Stone [67] and Burman [29] for the estimation of density on histograms and also by Arlot and Massart [10] and Arlot [7], [5] in the regression framework. Combining Lemma 1 and 2 of Barron and Sheu [15] we can show that

$$\frac{1}{2}e^{-3\left\|\ln\left(\frac{s_*}{s_M}\right)\right\|_\infty}\int_{\mathcal{Z}}\frac{\left(s_M - s_*\right)^2}{s_*}d\mu \leq \ell(s_*, s_M)$$

and thus assuming (**Abd**) we get

$$\frac{A_{\min}^3}{2A_*^4}\int_{\mathcal{Z}}\left(s_M - s_*\right)^2 d\mu \leq \ell\left(s_*, s_M\right) \ .$$

Now, since in the case of histograms the Kullback-leibler projection s_M is also the $L_2\left(\mu\right)$ projection of s_* onto M, we can apply Lemma 8.19 in Section 8.10 of Arlot [4] to show that assumption (**Ap**) is satisfied for $\beta_- = 1+\alpha^{-1}$, in the case where $\mathcal{Z}$ is the unit interval, $\mu = \text{Leb}$ is the Lebesgue measure on the unit interval, the partitions Λ_M are regular and the density s_* is a non-constant α-hölderian function.

5.4 Two directions of generalization

We present here two possible generalizations of the results exposed in Section 5.3. Models of piecewise constant densities have the particular property of been exponential models as well as the subset of positive functions in an affine space and we expose below strategies to extend our results in these two directions.
We first notice that the proofs of Theorems 5.3 and 5.4 of model selection follow from straightforward adaptations of the proofs of Theorem 2 and 3 in Arlot and Massart [10], only using the results given in Theorems 5.1, 5.2 and Lemmas 5.6 and 5.7 of Section 5.5.4 where the quantities of interest can be defined for more general models than histograms. For this reason, the proofs given in Arlot and Massart [10] give some general algebra to derive the properties of the slope heuristics considering a small collection of models and the main task is thus to deal with some fixed model. Theorems 5.1 and 5.2 respectively provide with a sharp control of the excess risk and the empirical excess risk for models of dimension not too large and not too small, and a control of the same quantities for models of small dimension. In Lemma 5.6 we derive a sharp control of the empirical excess risk in mean for models of reasonable dimension and in Lemma 5.7 we bound the difference between the bias and its empirical counterpart.
In the following, we emphasize on generalizations of Theorem 5.1. In fact, Lemma 5.7 that follows from Bernstein inequality can be easily extended to more general models and Lemma 5.6 is a straightforward corollary of Theorem 5.1. Moreover, Theorem 5.2 directly follows from the convergence in sup-norm of maximum likelihood estimators at the rate $\sqrt{D_M \ln\left(n\right)/n}$ as derived in the case of histograms in Proposition 5.2.

5.4.1 Affine spaces

We intend to point out here that results of Theorem 5.1 may be extended to more general linear models M than piecewise constant functions. Let us set

$$M = \left\{ s = \sum_{k=1}^{D_M} \beta_k \varphi_k \ ; \ \beta = \left(\beta_k\right)_{k=1}^{D_M} \in \mathbb{R}^{D_M} \right\} \tag{5.50}$$

the vector space of dimension D_M spanned by the basis $\left(\varphi_k\right)_{k=1}^{D_M}$ that we assume to be orthonormal in $L_2\left(P\right)$. We also set the subset $\widetilde{M}$ of the functions in M that are densities with respect to μ,

$$\widetilde{M} = \left\{ s \in M \ ; \ s \geq 0 \text{ and } \int_{\mathcal{Z}} s d\mu = 1 \right\} \ ,$$

and consider that the maximum likelihood estimator on $\widetilde{M}$ exists, denoted by $s_n\left(M\right)$.
The proof of Theorem 5.1, that we give in Section 5.5, relies on purpose on more general arguments than the ones strictly needed in the case of histograms. More precisely, using explicit formula 5.7 and 5.10 for the Kullback-Leibler projection and the histogram estimator,

we could have avoid the use of the slices in excess risk defined in (5.87) and (5.88) by controlling the excess risk and the empirical excess risk directly on the estimator. But our aim is to point out the generality of the method, and a careful look at the proof of Theorem 5.1 shows that for more general models as in (5.50), we achieve the same bounds for the excess risks (with different values of constants) if the five following points are satisfied :

- The target s_* is uniformly lower and upper bounded : for $A_{\min}, A_* > 0$,
$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_* (z) \leq \|s_*\|_\infty \leq A_* < +\infty$$
- The model is of reasonable dimension : $A_- (\ln n)^2 \leq D_M \leq A_+ \frac{n}{(\ln n)^2} \leq n$.
- $(\varphi_k)_{k=1}^{D_M}$ is a localized orthonormal basis in $(M, L_2(P))$: for some $r_M > 0$,
$$\left\| \sum_{k=1}^{D_M} \beta_k \varphi_k \right\|_\infty \leq r_M \sqrt{D_M} |\beta|_\infty \ . \tag{5.51}$$
- The Kullback-Leibler projection s_M is well-defined and the excess risk is, locally around s_M, close to the weighted $L_2(P)$ norm : positive constants A_H and L_H exist such that, if $\|s - s_M\|_\infty \leq \delta \leq A_H$ then
$$\left(\frac{1}{2} - L_H \delta\right) \left\| \frac{s - s_M}{s_M} \right\|_2^2 \leq P(Ks - Ks_M) \leq \left(\frac{1}{2} + L_H \delta\right) \left\| \frac{s - s_M}{s_M} \right\|_2^2 \ . \tag{5.52}$$
- The maximum likelihood estimator is consistent towards the Kullback-Leibler projection s_M at the rate $\sqrt{D_M \ln(n)/n}$: for any $\alpha > 0$, positive constants A_c and L_c exist such that
$$\mathbb{P}\left[\|s_n(M) - s_M\|_\infty \geq A_c \sqrt{\frac{D_M \ln n}{n}} \right] \leq L_c n^{-\alpha} \ . \tag{5.53}$$

Note that the assumption of lower regularity of the partition of Theorem 5.1 in the case of histograms, stating that $D_M \inf_{I \in \Lambda_M} \mu(I) \geq A_\Lambda > 0$ for some $A_\Lambda > 0$, is replaced here by the more general assumption of localized basis (5.51). It is easy to see using Lemma 5.1 that the two properties are equivalent in the case of histograms. Moreover, Property (5.52) is based in the case of histograms on the Pythagorean-like identity (5.9) given in Proposition 5.1 and remains a work in progress for more general models $\widetilde{M}$. In Csiszár and Matúš [35], general conditions are given under which Pythagorean-like identities for the Kullback-Leibler divergence hold true. In their terminology, the Kullback-Leibler projection is called "reverse I-projection". Among other results, they show Pythagorean-like identities in the context of convex sets, a property that is satisfied for $\widetilde{M}$, but considering the "I-projection" rather than the "reverse I-projection". Nevertheless, generalized reverse I-projections onto convex sets of probability measures can be found in Barron [12]. Property (5.53) remains an open issue for general linear models as well.

5.4.2 Exponential models

In this section, we briefly describe how our strategy of proofs, based on the notion of regular contrast, can be adapted to derive sharp bounds for the excess risks in the case of exponential models and possibly recover the slope heuristics in good cases. This work is still in progress. Let us set
$$M = \left\{ t = \sum_{k=1}^{D_M} \beta_k \varphi_k \ ; \ \beta = (\beta_k)_{k=1}^{D_M} \in \mathbb{R}^{D_M} \right\}$$

the linear vector space of dimension D_M spanned by the basis $(\varphi_k)_{k=1}^{D_M}$, that we assume to be orthonormal in $L_2(P)$. We assume that the constant function $\mathbf{1} \in M$ and that $M \subset L_\infty(\mu)$. Then we set the associated exponential model $\widetilde{M}$, defined to be

$$\widetilde{M} = \left\{ s = \exp(t) \ ; t \in M \text{ and } \int_{\mathcal{Z}} s d\mu = 1 \right\}$$

and consider the maximum likelihood estimator $s_n\left(\widetilde{M}\right)$ on $\widetilde{M}$. It is well-known (see for example Barron and Sheu [15] and also Csiszár and Matúš [35]) that in this case $s_n(M)$ exists with high probability as a solution of a family of linear constraints, and its uniqueness is a familiar consequence of the strict convexity of the log-likelihood. It is also well-known (see Lemma 3 of Barron and Sheu [15]) that the unknown density s_* has a unique Kullback-Leibler projection $s_{\widetilde{M}}$on $\widetilde{M}$, characterized by the following Pythagorean-like identity,

$$\mathcal{K}(s_*, s) = \mathcal{K}\left(s_*, s_{\widetilde{M}}\right) + \mathcal{K}\left(s_{\widetilde{M}}, s\right) \ .$$

This property is essential, as it follows that the excess risk on $\widetilde{M}$ is the Kullback-Leibler divergence with respect to the Kullback-Leibler projection $s_{\widetilde{M}}$,

$$P\left(Ks - Ks_{\widetilde{M}}\right) = \mathcal{K}\left(s_{\widetilde{M}}, s\right)$$

and by consequence, we can relate the excess risk on $\widetilde{M}$ to the $L_2(P)$ norm in M, due to the following lemma of Barron and Sheu [15].

Lemma 5.5 (Lemma 3, [15]) *Let p and q be two probability density functions with respect to μ such that $\|\ln(p/q)\|_\infty$ is finite. Then it holds*

$$\mathcal{K}(p, q) \geq \frac{1}{2} e^{-\|\ln(p/q)\|_\infty} \int p \left(\ln \frac{p}{q}\right)^2 d\mu$$

and

$$\mathcal{K}(p, q) \leq \frac{1}{2} e^{\|\ln(p/q) - c\|_\infty} \int p \left(\ln \frac{p}{q} - c\right)^2 d\mu \ ,$$

where c is any constant.

Hence, we have for any $s \in \widetilde{M}$,

$$0 < \frac{1}{2} e^{-\left\|\ln\left(s/s_{\widetilde{M}}\right)\right\|_\infty} \int s_{\widetilde{M}} \left(\ln \frac{s}{s_{\widetilde{M}}}\right)^2 d\mu \leq P\left(Ks - Ks_{\widetilde{M}}\right) \tag{5.54}$$

$$\leq \frac{1}{2} e^{\left\|\ln\left(s/s_{\widetilde{M}}\right)\right\|_\infty} \int s_{\widetilde{M}} \left(\ln \frac{s}{s_{\widetilde{M}}}\right)^2 d\mu \ . \tag{5.55}$$

Now, if we can show that

$$\left\| \ln \left(\frac{s_n\left(\widetilde{M}\right)}{s_{\widetilde{M}}} \right) \right\|_\infty \leq \frac{A_{cons}}{\sqrt{\ln n}}$$

for some positive constant A_{cons} and for all n sufficiently large, we can restrict our study to the subset of functions in $\widetilde{M}$ satisfying $\left\|\ln\left(s/s_{\widetilde{M}}\right)\right\|_\infty \leq \frac{A_{cons}}{\sqrt{\ln n}}$ - by the same type of arguments that are given in Section 7.3 of Chapter 7 - and so we have on this subset of interest, by inequalities (5.54) and (5.55),

$$\begin{aligned} P\left(Ks - Ks_{\widetilde{M}}\right) &\sim \frac{1}{2} \int s_{\widetilde{M}} \left(\ln \frac{s}{s_{\widetilde{M}}}\right)^2 d\mu \\ &= \frac{1}{2} \left\| \ln \frac{s}{s_{\widetilde{M}}} \right\|_2^2 + \frac{1}{2} \int \left(s_{\widetilde{M}} - s_*\right) \left(\ln \frac{s}{s_{\widetilde{M}}}\right)^2 d\mu \ . \end{aligned}$$

Moreover, for the right-hand term in the latter identity, it holds

$$\left|\frac{1}{2}\int \left(s_{\widetilde{M}}-s_*\right)\left(\ln\frac{s}{s_{\widetilde{M}}}\right)^2 d\mu\right| \leq \left\|s_{\widetilde{M}}-s_*\right\|_\infty \frac{1}{2}\left\|\ln\frac{s}{s_{\widetilde{M}}}\right\|^2_{L_2(\mu)}$$

which should be negligible in front of the weighted $L_2(P)$ norm $\frac{1}{2}\left\|\ln\frac{s}{s_{\widetilde{M}}}\right\|_2^2$ if the considered model $\widetilde{M}$ has a small bias in sup-norm and if the unknown density is uniformly bounded away from zero, in order to upper bound $\|\cdot\|_{L_2(\mu)}$ by $\|\cdot\|_{L_2(P)}$. Under the right assumptions on the smoothness of the target s_* and a suitable choice of M the assumption on the bias of the model should be satisfied if at least its dimension is not too small (a power of $\ln n$ should be again sufficient in many cases). The importance of a control in sup-norm for the bias of the models in maximum likelihood estimation of density has been pointed out by Stone [68] considering log-splines models. The author provides with a sharp control of the bias in sup-norm in this case, a work that should be inspiring for other situations and also in order to prove the consistency in sup-norm of $\ln\left(\frac{s_n(\widetilde{M})}{s_{\widetilde{M}}}\right)$. By consequence, we can conjecture that under reasonable assumptions, the weighted $L_2(P)$ norm described above is a good approximation of the excess risk on $\widetilde{M}$ for a model M of dimension not too small and it has the convenient property to be Hilbertian : on a subset of interest on $\widetilde{M}$,

$$P\left(Ks-Ks_{\widetilde{M}}\right) \sim \frac{1}{2}\left\|\ln\frac{s}{s_{\widetilde{M}}}\right\|_2^2 = \frac{1}{2}\left\|\ln s-\ln s_{\widetilde{M}}\right\|_2^2 \tag{5.56}$$

where $\ln(s)$ and $\ln\left(s_{\widetilde{M}}\right)$ belong to M.
Let us explain now how to take advantage of (5.56) for exponential models. The arguments given below are close in the spirit to arguments of Chapter 7, considering the log-linearity of exponential models, or in other words the linearity of the contrasted functions. If we set

$$t_M = \ln s_{\widetilde{M}} \in M$$

and for any $r \geq 0$,

$$\xi_n(r) = \mathbb{E}\left[\sup_{\substack{t\in M,\ \|t-t_M\|_2^2=2r \\ \int \exp(t)d\mu=1}} \left|(P-P_n)(t-t_M)\right|\right],$$

then, as claimed in Chapter 7, we can approximately write for models of reasonable dimension,

$$P\left(Ks_n\left(\widetilde{M}\right)-Ks_M\right) \sim \arg\max_{R_{n,D_M}\geq r\geq 0}\left\{\mathbb{E}\left[\sup_{s\in\widetilde{M},\ P\left(Ks-Ks_{\widetilde{M}}\right)=r}\left|(P-P_n)\left(Ks-Ks_{\widetilde{M}}\right)\right|\right]-r\right\}$$

where we assume that

$$P\left(Ks_n\left(\widetilde{M}\right)-Ks_{\widetilde{M}}\right) \leq \left\|\ln\left(\frac{s_n\left(\widetilde{M}\right)}{s_{\widetilde{M}}}\right)\right\|_\infty \leq R_{n,D_M} \leq \frac{A_{cons}}{\sqrt{\ln n}}$$

with high probability (of order $1-Ln^{-\alpha}$, $\alpha>0$). Then, from (5.56) we have for $R_{n,D_M}\geq r\geq 0$,

$$\mathbb{E}\left[\sup_{s\in\widetilde{M},\ P\left(Ks-Ks_{\widetilde{M}}\right)=r}\left|(P-P_n)\left(Ks-Ks_{\widetilde{M}}\right)\right|\right] \sim \xi_n(r)$$

and so

$$P\left(Ks_n\left(\widetilde{M}\right) - Ks_M\right) \sim \arg \max_{R_{n,D_M} \geq r \geq 0} \{\xi_n(r) - r\} \ . \tag{5.57}$$

By the same type of reasoning, we can also conjecture that for models of reasonable dimensions,

$$P_n\left(Ks_M - Ks_n\left(\widetilde{M}\right)\right) \sim \max_{R_{n,D_M} \geq r \geq 0} \{\xi_n(r) - r\} \ . \tag{5.58}$$

Moreover, in good cases satisfying assumptions of Corollary 7.2 we have

$$\xi_n(r) \sim \mathbb{E}^{1/2}\left[\left(\sup_{\substack{t \in M,\ \|t - t_M\|_2^2 = 2r \\ \int \exp(t) d\mu = 1}} |(P - P_n)(t - t_M)|\right)^2\right] \tag{5.59}$$

and if we define

$$t_{CS} = \sqrt{2r}\frac{\sum_{k=1}^{D_M}(P - P_n)(\varphi_k)\varphi_k}{\sqrt{\sum_{k=1}^{D_M}(P - P_n)^2(\varphi_k)}} + t_M \ ,$$

then it holds $\|t_{CS} - t_M\|_2^2 = 2r$ and

$$\begin{aligned} \sup_{t \in M,\ \|t - t_M\|_2^2 = 2r} |(P - P_n)(t - t_M)| &= (P - P_n)(t_{CS} - t_M) \\ &= \sqrt{2r}\sqrt{\sum_{k=1}^{D_M}(P - P_n)^2(\varphi_k)} \ . \end{aligned} \tag{5.60}$$

Now, assuming that $1 \gg R_{n,D_M} \geq L\sqrt{\frac{D_M \ln n}{n}}$ for a positive constant L sufficiently large, if we can prove that with high probability,

$$\|t_{CS} - t_M\|_\infty \leq R_{n,D_M} \text{ for } r \leq R_{n,D_M} \ ,$$

which is typically the case when $(\varphi_k)_{k=1}^{D_M}$ is a localized basis, then

$$\begin{aligned} \int \exp(t_{CS}) d\mu &\approx \int \exp(t_M) d\mu + \int (t_{CS} - t_M) d\mu \\ &\approx 1 \ . \end{aligned} \tag{5.61}$$

Finally, taking into account (5.59), (5.60) and (5.61), we can conjecture that under some assumptions on the model M that allow to control the sup-norm in a sufficiently sharp way, we would have

$$\xi_n(r) \sim \sqrt{\frac{2r}{n}\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}$$

for $r \leq R_{n,D_M}$ and so, using (5.57) and (5.58), as

$$\arg \max_{R_{n,D_M} \geq r \geq 0}\left\{\sqrt{\frac{2r}{n}\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)} - r\right\} = \max_{R_{n,D_M} \geq r \geq 0}\left\{\sqrt{\frac{2r}{n}\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)} - r\right\} = \frac{\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}{2n}$$

for $R_{n,D_M} \geq \sqrt{\frac{\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}{2n}}$, this would lead to

$$P\left(Ks_n\left(\widetilde{M}\right) - Ks_M\right) \sim P_n\left(Ks_M - Ks_n\left(\widetilde{M}\right)\right) \sim \frac{\sum_{k=1}^{D_M}\mathrm{Var}(\varphi_k)}{2n}$$

for models of reasonable dimensions having good enough properties with respect to the sup-norm.

5.5 Proofs

5.5.1 Proofs of Section 5.2

Proof of Lemma 5.1. Remind that, for all $I \in \Lambda_M$,

$$\varphi_I = (P(I))^{-1/2} \mathbf{1}_I \ .$$

Hence, $(\varphi_I)_{I\in\Lambda_M}$ is an orthonormal basis of $(M, L_2(P))$. Moreover, by (5.12) we have, for all $I \in \Lambda_M$,

$$P(I) \geq A_{\min}\mu(I) \geq A_{\min}A_\Lambda D_M^{-1} > 0$$

and so, by setting $r_M = (A_{\min}A_\Lambda)^{-1/2}$, we get for all $I \in \Lambda_M$,

$$(P(I))^{-1/2} \leq \sqrt{\frac{D_M}{A_{\min}A_\Lambda}} = r_M\sqrt{D_M} \ .$$

Now, as the elements φ_I for $I \in \Lambda_M$ have disjoint supports, we deduce that, for all $\beta = (\beta_I)_{I\in\Lambda_M} \in \mathbb{R}^{D_M}$,

$$\begin{aligned}\left\|\sum_{I\in\Lambda_M}\beta_I\varphi_I\right\|_\infty &\leq \max_{I\in\Lambda_M}\left\{|\beta_I|\,\|\varphi_I\|_\infty\right\}\\ &\leq \max_{I\in\Lambda_M}\left\{|\beta_I|\,(P(I))^{-1/2}\right\}\\ &\leq r_M\sqrt{D_M}\,|\beta|_\infty\end{aligned}$$

and Inequality (5.13) is then proved. Next, Inequality (5.14) easy follows by observing that, for any $s = \sum_{I\in\Lambda_M}\beta_I\varphi_I \in M$ satisfying $\|s\|_2 \leq 1$, we have

$$\max_{I\in\Lambda_M}|\beta_I| \leq \sqrt{\sum_{I\in\Lambda_M}\beta_I^2} \leq 1$$

and so

$$\|s\|_\infty \leq r_M\sqrt{D_M} \ .$$

■

Proof of Lemma 5.2. By (5.15)and (2.28), we have

$$\inf_{z\in\mathcal{Z}} s_M(z) \geq A_{\min} > 0 \ ,$$

then $\psi_{1,M}(z)$ and $(Ks_M)(z) = -\ln(s_M(z))$ are well defined for all $z \in \mathcal{Z}$. Moreover, as we assume $\|s - s_M\|_\infty < A_{\min}$, we have

$$\inf_{z\in\mathcal{Z}} s(z) = \inf_{z\in\mathcal{Z}}\{s_M(z) + (s - s_M)(z)\} \geq \inf_{z\in\mathcal{Z}} s_M(z) - \|s - s_M\|_\infty > 0$$

and

$$\left\|\frac{s - s_M}{s_M}\right\|_\infty \leq \frac{\|s - s_M\|_\infty}{A_{\min}} < 1$$

thus $(Ks)(z) = -\ln(s(z))$ is well defined for each $z \in \mathcal{Z}$ as well as $(s_M(z))^{-1}$ and $\ln\left(1 + \frac{s-s_M}{s_M}(z)\right)$, so the expansion (5.17) is a simple rewriting of the identity

$$(Ks)(z) - (Ks_M)(z) = -\ln\left(\frac{s(z)}{s_M(z)}\right) \ .$$

■

Proof of Lemma 5.3. Lemma **5.3** is straightforward, since

$$\psi_2'(x) = \frac{x}{1+x},\ x \in (-1, +\infty)\ .$$

Hence, for all $x \in \left[-\frac{\delta}{A_{\min}}, \frac{\delta}{A_{\min}}\right]$, with $0 \leq \delta \leq A_{\min}/2$,

$$\left|h'(x)\right| \leq \frac{\delta/A_{\min}}{1-\delta/A_{\min}} \leq 2\frac{\delta}{A_{\min}}\ ,$$

which yields the result. ■

Proof of Lemma 5.4. For $s \in M$ such that $\|s - s_M\|_\infty \leq \delta \leq \frac{A_{\min}}{2}$, we have

$$\inf_{z\in\mathcal{Z}} s(z) \geq \inf_{z\in\mathcal{Z}} s_M(z) - \|s - s_M\|_\infty \geq \frac{A_{\min}}{2} > 0 \text{ and } \left\|\frac{s-s_M}{s_M}\right\|_\infty \leq \frac{1}{2}\ .$$

and so, if $\int_{\mathcal{Z}} s d\mu = 1$ then $s \in \widetilde{M}$. Moreover, in this case, by (5.11) we have

$$P(Ks - Ks_M) = \mathcal{K}(s_M, s)$$

and it holds

$$\begin{aligned}
\mathcal{K}(s_M, s) &= \int_{\mathcal{Z}} \ln\left(\frac{s_M}{s}\right) s_M d\mu \\
&= \int_{\mathcal{Z}} -\ln\left(1 + \frac{s-s_M}{s_M}\right) s_M d\mu \\
&= \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \int_{\mathcal{Z}} \left(\frac{s-s_M}{s_M}\right)^k s_M d\mu \\
&= \int_{\mathcal{Z}} (s_M - s)\, d\mu + \frac{1}{2}\int_{\mathcal{Z}} \left(\frac{s-s_M}{s_M}\right)^2 s_M d\mu + \sum_{k=3}^{\infty} \frac{(-1)^k}{k} \int_{\mathcal{Z}} \left(\frac{s-s_M}{s_M}\right)^k s_M d\mu.
\end{aligned} \tag{5.62}$$

Now, as $\int_{\mathcal{Z}} s d\mu = 1$, we have

$$\int_{\mathcal{Z}} (s_M - s)\, d\mu = 0\ . \tag{5.63}$$

Moreover, notice that by (5.7), for all $I \in \Lambda_M$,

$$\int_{\mathcal{Z}} \mathbf{1}_I s_M d\mu = \frac{P(I)}{\mu(I)}\mu(I) = P(I) = \int_{\mathcal{Z}} \mathbf{1}_I s_* d\mu$$

and so, for all $t \in M$,

$$\int_{\mathcal{Z}} t \cdot s_M d\mu = \int_{\mathcal{Z}} t \cdot s_* d\mu\ .$$

Now, using the fact that $\left(\frac{s-s_M}{s_M}\right)^2 \in M$, it holds

$$\begin{aligned}
\frac{1}{2}\int_{\mathcal{Z}} \left(\frac{s-s_M}{s_M}\right)^2 s_M d\mu &= \frac{1}{2}\int_{\mathcal{Z}} \left(\frac{s-s_M}{s_M}\right)^2 s_* d\mu \\
&= \frac{1}{2} P\left(\frac{s-s_M}{s_M}\right)^2 \\
&= \frac{1}{2}\left\|\frac{s-s_M}{s_M}\right\|_2^2\ .
\end{aligned} \tag{5.64}$$

Moreover, we have

$$\begin{aligned}
&\left|\sum_{k=3}^{\infty}\frac{(-1)^k}{k}\int_{\mathcal{Z}}\left(\frac{s-s_M}{s_M}\right)^k s_M d\mu\right| \\
&\leq \frac{1}{3}\int_{\mathcal{Z}}\left(\frac{s-s_M}{s_M}\right)^2 s_M d\mu \times \sum_{j=1}^{\infty}\left\|\frac{s-s_M}{s_M}\right\|_{\infty}^{j} \\
&= \frac{1}{3}\int_{\mathcal{Z}}\left(\frac{s-s_M}{s_M}\right)^2 s_* d\mu \times \sum_{j=1}^{\infty}\left\|\frac{s-s_M}{s_M}\right\|_{\infty}^{j} \\
&\leq \left\|\frac{s-s_M}{s_M}\right\|_2^2 \frac{2\delta}{3A_{\min}} \ . \qquad (5.65)
\end{aligned}$$

Inequality (5.21) then follows by using (5.63), (5.64) and (5.65) in (5.62). ■

5.5.2 Proof of Section 5.3.1

Proof of Proposition 5.2. Let $\beta > 0$ to be fixed later. Recall that, by (5.7) and (5.10),

$$\begin{aligned}
s_n(M) &= \sum_{I\in\Lambda_M}\frac{P_n(I)}{\mu(I)}\mathbf{1}_I \ , \\
s_M &= \sum_{I\in\Lambda_M}\frac{P(I)}{\mu(I)}\mathbf{1}_I \ .
\end{aligned}$$

Hence, the sup-norm of the difference can be written

$$\|s_n(M) - s_M\|_{\infty} = \sup_{I\in\Lambda_M}\frac{|(P_n - P)(I)|}{\mu(I)} \ . \qquad (5.66)$$

By Bernstein's inequality (7.46) applied for the random variable $\mathbf{1}_{\xi\in I}$ we get, for all $x > 0$,

$$\mathbb{P}\left[|(P_n - P)(I)| \geq \sqrt{\frac{2P(I)\,x}{n}} + \frac{x}{3n}\right] \leq 2\exp(-x) \ .$$

Taking $x = \beta \ln n$ and normalizing by the quantity $\mu(I) > 0$ we get

$$\mathbb{P}\left[\frac{|(P_n - P)(I)|}{\mu(I)} \geq \frac{1}{\mu(I)}\sqrt{\frac{2\beta P(I)\ln n}{n}} + \frac{\beta\ln n}{\mu(I)\,3n}\right] \leq 2n^{-\beta} \ . \qquad (5.67)$$

Now, by (5.22) and (5.23),

$$0 < \frac{1}{\mu(I)} \leq \frac{D_M}{A_\Lambda} \qquad (5.68)$$

and

$$\frac{\sqrt{P(I)}}{\mu(I)} \leq \sqrt{\frac{A_*}{\mu(I)}} \leq \sqrt{\frac{A_* D_M}{A_\Lambda}} \ . \qquad (5.69)$$

So, injecting (5.68) and (5.69) in (5.67) and using the fact that $D_M \leq A_+ \frac{n}{(\ln n)^2}$ we get

$$\mathbb{P}\left[\frac{|(P_n - P)(I)|}{\mu(I)} \geq A_c\sqrt{\frac{D_M \ln n}{n}}\right] \leq 2n^{-\beta} \ , \qquad (5.70)$$

where $A_c = \max\left\{\sqrt{\frac{2\beta A_*}{A_\Lambda}}\ ;\ \frac{\beta\sqrt{A_+}}{3A_\Lambda}\right\}$. We then deduce from (5.66) and (5.70) that

$$\mathbb{P}\left[\left\|s_n\left(M\right)-s_M\right\|_\infty \geq A_c\sqrt{\frac{D_M\ln n}{n}}\right] \leq \frac{2D_M}{n^\beta}$$

and, since $D_M \leq n$, taking $\beta = \alpha+1$ yields Inequality (5.24). ■

Proof of Proposition 5.3. Let $\beta > 0$ to be fixed later. Recall that, by (5.7) and (5.10),

$$s_n\left(M\right) = \sum_{I\in\Lambda_M}\frac{P_n\left(I\right)}{\mu\left(I\right)}\mathbf{1}_I\ , \tag{5.71}$$

$$s_M = \sum_{I\in\Lambda_M}\frac{P\left(I\right)}{\mu\left(I\right)}\mathbf{1}_I\ . \tag{5.72}$$

Hence, by (5.25) and (5.72) we get $\inf s_M\left(z\right) \geq A_{\min} > 0$. By (5.71) and (5.72) we have

$$\left\|\frac{s_n\left(M\right)-s_M}{s_M}\right\|_\infty = \sup_{I\in\Lambda_M}\frac{\left|\left(P_n-P\right)\left(I\right)\right|}{P\left(I\right)}\ . \tag{5.73}$$

By Bernstein's inequality (7.46) applied for the random variable $\mathbf{1}_{\xi\in I}$ we get, for all $x > 0$,

$$\mathbb{P}\left[\left|\left(P_n-P\right)\left(I\right)\right| \geq \sqrt{\frac{2P\left(I\right)x}{n}}+\frac{x}{3n}\right] \leq 2\exp\left(-x\right)\ .$$

Taking $x = \beta\ln n$ and normalizing by the quantity $P\left(I\right) \geq A_{\min}\mu\left(I\right) > 0$ we get

$$\mathbb{P}\left[\frac{\left|\left(P_n-P\right)\left(I\right)\right|}{P\left(I\right)} \geq \sqrt{\frac{2\beta\ln n}{P\left(I\right)n}}+\frac{\beta\ln n}{P\left(I\right)3n}\right] \leq 2n^{-\beta}\ . \tag{5.74}$$

Now, by (5.25) and (5.26), we have

$$0 < \frac{1}{P\left(I\right)} \leq \frac{D_M}{A_{\min}A_\Lambda}\ . \tag{5.75}$$

Hence, using (5.75) in (5.74) and using the fact that $D_M \leq A_+\frac{n}{(\ln n)^2}$ we get

$$\mathbb{P}\left[\frac{\left|\left(P_n-P\right)\left(I\right)\right|}{P\left(I\right)} \geq A_c\sqrt{\frac{D_M\ln n}{n}}\right] \leq 2n^{-\beta}\ , \tag{5.76}$$

where $A_c = \max\left\{\sqrt{\frac{2\beta}{A_\Lambda A_{\min}}}\ ;\ \frac{\beta\sqrt{A_+}}{3A_{\min}A_\Lambda}\right\}$. We then deduce from (5.73) and (5.76) that

$$\mathbb{P}\left[\left\|s_n\left(M\right)-s_M\right\|_\infty \geq A_c\sqrt{\frac{D_M\ln n}{n}}\right] \leq \frac{2D_M}{n^\beta}$$

and, since $D_M \leq n$, taking $\beta = \alpha+1$ yields Inequality (5.27). ■

5.5.3 Proofs of Theorems 5.1 and 5.2

In order to introduce the quantities of interest, we recall some notations stated below and add some new definitions. As usual, M denotes the finite dimensional linear vector space of piecewise constant functions with respect to the finite partition Λ_M. Moreover, we write $D_M = |\Lambda_M|$ the linear dimension of M. Assuming (5.46) and (5.45) we have, for all $I \in \Lambda_M$, $P(I) > 0$ and so, if we set

$$\varphi_I = \frac{\mathbf{1}_I}{\sqrt{P(I)}} , \qquad I \in \Lambda_M ,$$

the family $(\varphi_I)_{I\in\Lambda_M}$ is an orthonormal basis of $(M, L_2(P))$. We set

$$\nu_n = \max\left\{\sqrt{\frac{\ln n}{D_M}}, \sqrt{\frac{D_M \ln n}{n}}\right\} . \tag{5.77}$$

In what follows $\alpha > 0$ is fixed and for some positive constant A_∞ to be chosen in the proof of Theorem 5.1 and satisfying

$$A_\infty \geq A_c > 0$$

where A_c is defined in Proposition 5.2 and only depends on A_Λ, A_*, A_+ and α, we set

$$\tilde{R}_{n,D_M,\alpha} = A_\infty \sqrt{\frac{D_M \ln n}{n}} \tag{5.78}$$

and

$$\Omega_{\infty,\alpha} = \left\{\|s_n(M) - s_M\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\} .$$

By Proposition 5.2 it holds, since $A_\infty \geq A_c$,

$$\mathbb{P}\left[\Omega^c_{\infty,\alpha}\right] \leq 2n^{-\alpha} . \tag{5.79}$$

Moreover, our analysis is localized on the subset

$$B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) = \left\{s \in M, \|s - s_M\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\} .$$

Assuming that

$$D_M \leq A_+ n (\ln n)^{-2}$$

we have, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\tilde{R}_{n,D_M,\alpha} \leq \frac{A_{\min}}{2} \tag{5.80}$$

where $A_{\min}$ is defined in (5.45). Now, assuming (5.45), we have by (5.80) and Lemma 5.2, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$, for every $s \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ and all $z \in \mathcal{Z}$,

$$(Ks)(z) - (Ks_M)(z) = \psi_{1,M}(z)(s - s_M)(z) + \psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right) \tag{5.81}$$

where

$$\psi_{1,M}(z) = -\frac{1}{s_M(z)}$$

and, for all $t \in (-1, +\infty)$,

$$\psi_2(t) = t - \ln(1 + t) .$$

Recall that, by (5.45),

$$\|\psi_{1,M}\|_\infty \leq \left(\inf_{z\in\mathcal{Z}} |s_M(z)|\right)^{-1} \leq A^{-1}_{\min} . \tag{5.82}$$

Moreover, by (5.80) and Lemma 5.3 we have, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$, for all $s \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ and all $z \in \mathcal{Z}$, using that $\psi_2(0) = 0$,

$$\left|\psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right)\right| \leq \left|\left(\frac{s - s_M}{s_M}\right)(z)\right| . \tag{5.83}$$

We also have by (5.80) and Lemma 5.3, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$, for every $s, t \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ and all $z \in \mathcal{Z}$,

$$\left|\psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right) - \psi_2\left(\left(\frac{t - s_M}{s_M}\right)(z)\right)\right| \leq 2A_{\min}^{-2}\tilde{R}_{n,D_M,\alpha}\left|(t - s)(z)\right| . \tag{5.84}$$

For convenience, we will use the following notation,

$$\psi_2 \circ \left(\frac{s - s_M}{s_M}\right) : z \in \mathcal{Z} \longmapsto \psi_2\left(\left(\frac{s - s_M}{s_M}\right)(z)\right) .$$

We now define slices of excess risk on the model $\widetilde{M}$. We set, for all $C > 0$,

$$\mathcal{F}_C = \left\{s \in \widetilde{M} \ ; \ \|\psi_{1,M} \cdot (s - s_M)\|_2^2 \leq 2C\right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) \tag{5.85}$$

$$\mathcal{F}_{>C} = \left\{s \in \widetilde{M} \ ; \ \|\psi_{1,M} \cdot (s - s_M)\|_2^2 > 2C\right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) \tag{5.86}$$

and for any interval J,

$$\mathcal{F}_J = \left\{s \in \widetilde{M} \ ; \ \frac{1}{2}\|\psi_{1,M} \cdot (s - s_M)\|_2^2 \in J\right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) . \tag{5.87}$$

We also define, for all $L \geq 0$,

$$D_L = \left\{s \in \widetilde{M} \ ; \ \|\psi_{1,M} \cdot (s - s_M)\|_2^2 = 2L\right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) . \tag{5.88}$$

By Lemma 5.4, we have, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$ and for any $s \in B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ such that $\int_{\mathcal{Z}} s d\mu = 1$,

$$0 < \left(\frac{1}{2} - \frac{2}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)\|\psi_{1,M} \cdot (s - s_M)\|_2^2 \leq \mathcal{K}(s_M, s) = P(Ks - Ks_M) \tag{5.89}$$

$$\leq \left(\frac{1}{2} + \frac{2}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)\|\psi_{1,M} \cdot (s - s_M)\|_2^2 . \tag{5.90}$$

Finally, notice that, if we assume (5.45) and (**Alr**), then by Proposition 5.1, if we set $r_M = (A_{\min}A_\Lambda)^{-1/2}$ then for all $z \in \mathcal{Z}$,

$$\sup_{s \in M, \ \|s\|_2 \leq 1} \|s\|_\infty \leq r_M\sqrt{D_M} \tag{5.91}$$

and moreover, for all $\beta = (\beta_I)_{I \in \Lambda_M} \in \mathbb{R}^{D_M}$,

$$\left\|\sum_{I \in \Lambda_M} \beta_I \varphi_I\right\|_\infty \leq r_M\sqrt{D_M}\,|\beta|_\infty . \tag{5.92}$$

Proofs of Theorems 5.1 and 5.2.

Proof of Theorem (3.22). We divide the proof of Theorem 5.1 in four parts corresponding to the four Inequalities (5.32), (5.33), (5.34) and (5.35). The values of A_0 and A_∞, respectively defined in (5.31) and (5.78), will then be fixed at the end of the proof. Note that, since $D_M \geq A_- (\ln n)^2$, we have $D_M \geq 2$ for all $n \geq n_0(A_-)$ so we can assume in the following that $D_M \geq 2$.

Proof of Inequality (5.32). By (5.78), it holds for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$1 - \frac{4}{3A_{\min}} \tilde{R}_{n,D_M,\alpha} > \frac{1}{2} \ .$$

Let $r \in (1,2]$ to be chosen later and $C, \widetilde{C} > 0$ such that

$$rC = \frac{D_M - 1}{2n} \tag{5.93}$$

and, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\widetilde{C} = \left(1 - \frac{4}{3A_{\min}} \tilde{R}_{n,D_M,\alpha}\right) C > 0 \ .$$

By inequality (5.89), if

$$P(Ks_n(M) - Ks_M) \leq \widetilde{C} \quad \text{and} \quad \|s_n(M) - s_M\|_\infty \leq \tilde{R}_{n,D_M,\alpha}$$

then

$$\|\psi_{1,M} \cdot (s_n(M) - s_M)\|_2^2 \leq 2C \ ,$$

for all $n \geq n_0(A_+, A_{\min}, A_\infty)$. Hence, by inequality (5.79), we get for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\begin{aligned} \mathbb{P}\left(P(Ks_n(M) - Ks_M) \leq \widetilde{C}\right) &\leq \mathbb{P}\left(\left\{P(Ks_n(M) - Ks_M) \leq \widetilde{C}\right\} \bigcap \Omega_{\infty,\alpha}\right) + 2n^{-\alpha} \\ &\leq \mathbb{P}\left(\left\{\|\psi_{1,M} \cdot (s_n(M) - s_M)\|_2^2 \leq 2C\right\} \bigcap \Omega_{\infty,\alpha}\right) + 2n^{-\alpha} \ . \end{aligned} \tag{5.94}$$

Now, by definition of the slices $\mathcal{F}_C$ and $\mathcal{F}_{>C}$ respectively given in (5.85) and (5.86), it holds

$$\begin{aligned} &\mathbb{P}\left(\left\{\|\psi_{1,M} \cdot (s_n(M) - s_M)\|_2^2 \leq 2C\right\} \bigcap \Omega_{\infty,\alpha}\right) \\ &\leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s \in \mathcal{F}_{>C}} P_n(Ks - Ks_M)\right) \\ &\leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s \in \mathcal{F}_{(C,rC]}} P_n(Ks - Ks_M)\right) \\ &= \mathbb{P}\left(\sup_{s \in \mathcal{F}_C} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_{(C,rC]}} P_n(Ks_M - Ks)\right) \ . \end{aligned} \tag{5.95}$$

Now, as by (5.93) we have

$$\frac{D_M}{8n} \leq C \leq (1 + A_4 \nu_n)^2 \frac{D_M - 1}{2n}$$

where A_4 is defined in Lemma 5.13, we can apply Lemma 5.13 with $\alpha = \beta$, $A_l = 1/8$ and it holds, for all $n \geq n_0 (A_-, A_+, A_{\min}, r_M, A_\infty, \alpha)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_C} P_n (Ks_M - Ks) \geq \left(1 + L_{A_-, A_{\min}, A_\infty, r_M, \alpha} \times \nu_n\right) \sqrt{\frac{2C (D_M - 1)}{n}} - C\right] \leq 2n^{-\alpha} , \tag{5.96}$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D_M}},\ \sqrt{\frac{D_M \ln n}{n}}\right\}$. Moreover, we can apply Lemma 5.15 with

$$\alpha = \beta,\ A_l = 1/8,\ A_u = 1/2$$

and

$$A_\infty \geq 32\sqrt{2} B_2 A_* r_M ,$$

and since $rC = (D_M - 1)/2n$, it gives, for all $n \geq n_0 (A_+, A_-, r_M, A_{\min}, A_\infty)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{(C, rC]}} P_n (Ks_M - Ks) \leq \left(\frac{1}{2} - L_{A_-, A_{\min}, A_\infty, \alpha} \times \nu_n\right) \frac{D_M - 1}{n}\right) \leq 2n^{-\alpha} , \tag{5.97}$$

Now, from (5.96) and (5.97) we can deduce that a positive constant $\tilde{A}_0$ exists, only depending on A_-, $A_{\min}$, A_∞, r_M and α, such that for all $n \geq n_0 (A_-, A_+, A_{\min}, r_M, A_\infty, \alpha)$, it holds on the same event of probability at least $1 - 4n^{-\alpha}$,

$$\begin{aligned} \sup_{s \in \mathcal{F}_C} P_n (Ks_M - Ks) &\leq \left(1 + \tilde{A}_0 \nu_n\right) \sqrt{\frac{2C (D_M - 1)}{n}} - C \\ &= \left(1 + \tilde{A}_0 \nu_n\right) \frac{D_M - 1}{n} \frac{1}{\sqrt{r}} - \frac{D_M - 1}{2n} \frac{1}{r} \end{aligned} \tag{5.98}$$

and

$$\sup_{s \in \mathcal{F}_{(C, rC]}} P_n (Ks_M - Ks) \geq \left(1 - 2\tilde{A}_0 \nu_n\right) \frac{D_M - 1}{2n} . \tag{5.99}$$

Hence, from (5.98) and (5.99) we can deduce, using (5.94) and (5.95), that if we choose $r \in (1, 2]$ such that

$$\left(1 - 2\tilde{A}_0 \nu_n\right) r - 2\left(1 + \tilde{A}_0 \nu_n\right) \sqrt{r} + 1 > 0 \tag{5.100}$$

then, for all $n \geq n_0 (A_-, A_+, A_{\min}, r_M, A_\infty, \alpha)$, $P (Ks_n (M) - Ks_M) \geq C$ with probability at least $1 - 6n^{-\alpha}$. Moreover, since

$$A_- (\ln n)^2 \leq D_M \leq A_+ n (\ln n)^{-2}$$

we have, for all $n \geq n_0 \left(A_+, A_-, \tilde{A}_0\right)$,

$$\tilde{A}_0 \nu_n \leq \frac{1}{4} \tag{5.101}$$

and so, for all $n \geq n_0 \left(A_+, A_-, \tilde{A}_0\right)$, simple computations using (5.101) show that by taking

$$r = 1 + 48\sqrt{\tilde{A}_0 \nu_n} \tag{5.102}$$

inequality (5.100) is satisfied. Notice that, for all $n \geq n_0 \left(A_+, A_-, \tilde{A}_0\right)$, $0 < 48\sqrt{\tilde{A}_0 \nu_n} < 1$, so that $r \in (1, 2]$. Finally, we can compute C by (3.107) and (3.118), for all $n \geq n_0 \left(A_+, A_-, \tilde{A}_0\right)$,

$$C = \frac{rC}{r} = \frac{1}{1 + 48\sqrt{\tilde{A}_0 \nu_n}} \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \geq \left(1 - 48\sqrt{\tilde{A}_0 \nu_n}\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 > 0 . \tag{5.103}$$

The result then follows the fact that by (5.103) and (5.77), it holds for all $n \geq n_0\left(A_+, A_-, A_\infty, A_{\min}, \tilde{A}_0\right)$,

$$\begin{aligned}
\widetilde{C} &= \left(1 - \frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right) C \\
&\geq \left(1 - \frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)\left(1 - 48\sqrt{\tilde{A}_0\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \\
&\geq \left(1 - \frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)\left(1 - 48\sqrt{\tilde{A}_0\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \\
&\geq \left(1 - L_{A_\infty,A_{\min}}\nu_n\right)\left(1 - 48\sqrt{\tilde{A}_0\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \\
&\geq \left(1 - L_{A_\infty,A_{\min},\tilde{A}_0}\sqrt{\nu_n}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \ ,
\end{aligned}$$

where the constant $\tilde{A}_0$ only depends on A_-, $A_{\min}$, A_∞, r_M and α. ■

To prove inequalities (5.33), (5.34), (5.35) and Theorem 5.2 it suffices to adapt the proofs of inequalities (3.23), (3.24), (3.25) and Theorem 3.2 given in Section 3.6.3 of Chapter 4 in the same way that we just did in the proof of inequality (5.32). We thus skip these proofs as they are now straightforward.

5.5.4 Proofs of Section 5.3.3

Given Lemmas 5.6 and 5.7 below, the proofs of Theorems 5.3 and 5.4 follow from straightforward adaptations of the proofs of Theorems 4.1 and 4.2 given in Section 4.4 of Chapter 4.

Lemma 5.6 *Let $A_{\mathcal{M},-} > 0$. Assume (**P2**), (**Abd**) and (**Alr**) of the set of assumptions defined in Section 5.3.3. Then for every model M of dimension D_M such that*

$$A_{\mathcal{M},-}\left(\ln n\right)^2 \leq D_M \leq A_{\mathcal{M},+} n \left(\ln n\right)^{-2} \ ,$$

we have, for all $n \geq n_0\left(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_\Lambda, A_{\min}, A_*, \alpha_{\mathcal{M}}\right)$,

$$\left(1 - L_{A_{\mathcal{M},+},A_{\mathcal{M},-},A_*,A_{\min},A_\Lambda}\varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n} \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \tag{5.104}$$

$$\leq \left(1 + L_{A_{\mathcal{M},+},A_{\mathcal{M},-},A_*,A_{\min},A_\Lambda}\varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n} \tag{5.105}$$

where $\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}, \left(\frac{D_M \ln n}{n}\right)^{1/4}\right\}$ is defined in Theorem 5.1.

Proof. Under assumptions of Lemma 5.6 we can apply Theorem 5.1 with $\alpha = 2 + \alpha_{\mathcal{M}}$. For all $n \geq n_0\left(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_{\min}, A_*, \alpha_{\mathcal{M}}\right)$, we thus have on an event $\Omega_1\left(M\right)$ of probability at least $1 - 6n^{-2-\alpha_{\mathcal{M}}}$,

$$\left(1 - \varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n} \leq P_n\left(Ks_M - Ks_n\left(M\right)\right) \leq \left(1 + \varepsilon_n^2\left(M\right)\right)\frac{D_M - 1}{2n} \tag{5.106}$$

where

$$\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}, \left(\frac{D_M \ln n}{n}\right)^{1/4}\right\} \geq A_0 n^{-1/8} \tag{5.107}$$

as $D_M \geq 1$. Moreover, we have,

$$\begin{aligned}
0 &\leq P_n\left(Ks_M - Ks_n(M)\right) = P_n\left(\ln\left(\frac{s_n(M)}{s_M}\right)\right) \\
&= P_n\left(\sum_{I\in\Lambda_M} \ln\left(\frac{P_n(I)}{P(I)}\right)\mathbf{1}_I\right) = \sum_{I\in\Lambda_M} \ln\left(P_n(I)\right)P_n(I) + \sum_{I\in\Lambda_M} \ln\left(\frac{1}{P(I)}\right)P_n(I) \\
&\leq \max_{I\in\Lambda_M}\left\{\ln\left(\frac{1}{P(I)}\right)\right\} \leq \ln\left(\left(A_{\min}A_\Lambda\right)^{-1}D_M\right) \ , \qquad (5.108)
\end{aligned}$$

where the last inequality follows from (**Abd**) and (**Alr**). We also have

$$\begin{aligned}
&\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] \\
&= \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{\Omega_1(M)}\right] + \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{(\Omega_1(M))^c}\right]. \qquad (5.109)
\end{aligned}$$

Hence, as $n \geq D_M \geq A_{\mathcal{M},-}\left(\ln n\right)^2$, it comes from (5.107) and (5.108) that, for all $n \geq n_0\left(A_{\mathcal{M},-}, A_0, A_{\min}, A_\Lambda\right)$,

$$0 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{(\Omega_1(M))^c}\right] \leq 6\ln\left(\left(A_{\min}A_\Lambda\right)^{-1}D_M\right)n^{-2-\alpha_{\mathcal{M}}} \leq \varepsilon_n^2(M)\frac{D_M-1}{2n} \qquad (5.110)$$

and, as we can see that $\varepsilon_n(M) < 1$ for all $n \geq n_0(A_0)$, we have by (5.106), for all $n \geq n_0\left(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_0, A_{\min}, A_*, \alpha_{\mathcal{M}}\right)$,

$$\begin{aligned}
\left(1-6n^{-2-\alpha_{\mathcal{M}}}\right)\left(1-\varepsilon_n^2(M)\right)\frac{D_M-1}{2n} &\leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{\Omega_1(M)}\right] \qquad (5.111) \\
&\leq \left(1-6n^{-2-\alpha_{\mathcal{M}}}\right)\left(1+\varepsilon_n^2(M)\right)\frac{D_M-1}{2n} \ . \qquad (5.112)
\end{aligned}$$

Finally, noticing that $n^{-2-\alpha_{\mathcal{M}}} \leq A_0^{-2}\varepsilon_n^2(M)$ by (5.107), we can use (5.110), (5.111) and (5.112) in (5.109) to conclude by straightforward computations that

$$L_{A_{\mathcal{M},+},A_{\mathcal{M},-},A_*,A_{\min},A_\Lambda} = 6A_0^{-2} + 2$$

is convenient in (5.104) and (5.105), as A_0 only depends on $\alpha_{\mathcal{M}}, A_-, A_+, A_*, A_{\min}$ and A_Λ. ■

Lemma 5.7 *Let $\alpha > 0$. Assume that (**Abd**) of Section 5.3.3 is satisfied. Then by setting $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$, we have for all $M \in \mathcal{M}_n$,*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \sqrt{\frac{4A_*\alpha\ell(s_*, s_M)\ln n}{A_{\min}n}} + \ln\left(\frac{A_*}{A_{\min}}\right)\frac{\alpha\ln n}{3n}\right) \leq 2n^{-\alpha} \qquad (5.113)$$

*and if moreover, assumptions (**P2**), (**Abd**) and (**Alr**) of Section 5.3.3 hold, then a positive constant A_d exists, depending only in A_*, $A_{\min}$ and α such that, for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},-}\left(\ln n\right)^2 \leq D_M$ and for all $n \geq n_0\left(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda\right)$,*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + A_d\frac{\ln n}{\sqrt{D_M}}\mathbb{E}\left[p_2(M)\right]\right) \leq 2n^{-\alpha} \ , \qquad (5.114)$$

where $p_2(M) = P_n\left(Ks_M - Ks_n(M)\right)$.

Proof. First recall that

$$s_M = \sum_{I \in \Lambda_M} \frac{P(I)}{\mu(I)} \mathbf{1}_I = \sum_{I \in \Lambda_M} \left(\int_I s_* \frac{d\mu}{\mu(I)} \right) \mathbf{1}_I \ .$$

Thus by (**Abd**) we deduce that

$$0 < A_{\min} \leq \inf_{z \in \mathcal{Z}} s_M(z) \leq \|s_M\|_\infty \leq A_* < +\infty \ . \tag{5.115}$$

Now, as we have

$$Ks_M - Ks_* = -\ln\left(\frac{s_M}{s_*}\right) \ ,$$

we get, by (**Abd**) and (5.115), that

$$\|Ks_M - Ks_*\|_\infty \leq \ln\left(\frac{A_*}{A_{\min}}\right) \ . \tag{5.116}$$

Hence, by Lemma 1 of Barron and Sheu [15], we have

$$P\left[(Ks_M - Ks_*)^2\right] \leq 2\exp\left(\|Ks_M - Ks_*\|_\infty\right) \mathcal{K}(s_*, s_M) \ .$$

By Proposition 5.1, we also have

$$\mathcal{K}(s_*, s_M) = P(Ks_M - Ks_*) = \ell(s_*, s_M)$$

and thus by (5.116), it holds

$$P\left[(Ks_M - Ks_*)^2\right] \leq \frac{2A_*}{A_{\min}} \ell(s_*, s_M) \ . \tag{5.117}$$

We are now ready to apply Bernstein's inequality (7.46) to

$$\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*) \ .$$

By (5.116) and (5.117) we have, for any $x > 0$,

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \sqrt{\frac{4A_* \ell(s_*, s_M)\, x}{A_{\min} n}} + \ln\left(\frac{A_*}{A_{\min}}\right) \frac{x}{3n} \right) \leq 2\exp(-x) \ .$$

Hence, taking $x = \alpha \ln n$ we have

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \sqrt{\frac{4A_* \alpha \ell(s_*, s_M) \ln n}{A_{\min} n}} + \ln\left(\frac{A_*}{A_{\min}}\right) \frac{\alpha \ln n}{3n} \right) \leq 2n^{-\alpha} \ , \tag{5.118}$$

which yields Inequality (5.113). Now, by noticing the fact that $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$, and by using it in (5.118) with $a = \ell(s_*, s_M)$, $b = \frac{A_* \alpha \ln n}{A_{\min} n}$ and $\eta = D_M^{-1/2}$, we obtain

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + \left(\frac{A_*}{A_{\min}} \sqrt{D_M} + \frac{1}{3} \ln\left(\frac{A_*}{A_{\min}}\right) \right) \frac{\alpha \ln n}{n} \right) \leq 2n^{-\alpha} \ . \tag{5.119}$$

Then, for a model M such that $A_{\mathcal{M},-} (\ln n)^2 \leq D_M \leq A_{\mathcal{M},+} n (\ln n)^{-2}$, we can apply Lemma 5.6 and by (5.104), it holds for all $n \geq n_0(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_\Lambda, A_{\min}, A_*, \alpha_{\mathcal{M}})$,

$$\left(1 - L_{A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda} \varepsilon_n^2(M)\right) \frac{D_M - 1}{2n} \leq \mathbb{E}\left[p_2(M)\right] \tag{5.120}$$

where $\varepsilon_n(M) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}, \left(\frac{D_M \ln n}{n}\right)^{1/4}\right\}$. Moreover as

$$A_{\mathcal{M},-}(\ln n)^2 \leq D_M \leq A_{\mathcal{M},+} n (\ln n)^{-2} \ ,$$

we can deduce that for all $n \geq n_0(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda)$,

$$L_{A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda} \varepsilon_n^2(M) \leq 1/2$$

and we have by (5.120), $\mathbb{E}[p_2(M)] \geq \frac{D_M}{8n}$ for all $n \geq n_0(A_{\mathcal{M},+}, A_{\mathcal{M},-}, A_*, A_{\min}, A_\Lambda)$. This allows, using (5.119), to conclude the proof by simple computations. ■

5.5.5 Technical lemmas

We state here some lemmas needed in the proofs of Theorem 5.1. Their proofs are quite similar to the proofs given in Section 7.5.1 of Chapter 3 as we use the same generic approach exposed in details in Chapter 7. More precisely, the least-squares contrast in regression and the Kullback-Leibler contrast satisfy the same formal property of expansion (5.81) and the models that we consider are endowed with localized basis. Moreover, the histograms estimators in MLE satisfy the assumption of consistency in sup-norm (**H5**) of Section 3.3.1 required in the case of regression, at the rate $R_{n,D_M,\alpha} \propto \sqrt{\frac{D_M \ln n}{n}}$. The main technical difference comes from the fact that the Kullback-Leibler excess risk is only close to an Hilbertian norm on the considered functions of $B_{(M,L_\infty)}(s_M, \tilde{R}_{n,D_M,\alpha})$, whereas in the least-squares regression the excess risk is the Hilbertian $L_2(P)$ norm itself.

Lemma 5.8 *Assume (5.45), (**Alr**) and $D_M \geq 2$. Then for any $\beta > 0$, a positive constant $L_{r_M,\beta}$ exists, such that by setting*

$$\tau_n = L_{r_M,\beta}\left(\sqrt{\frac{\ln n}{D_M}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right) \ ,$$

we have

$$\mathbb{P}\left[\sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2(\varphi_I)} \geq (1+\tau_n)\sqrt{\frac{D_M - 1}{n}}\right] \leq n^{-\beta} \ .$$

Proof. By Cauchy-Schwarz inequality we have

$$\chi_M = \sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2(\varphi_I)} = \sup_{s \in \mathcal{F}_{(C,rC]}} \{|(P_n - P)(s)| \ ; \ s \in M \ \& \ \|s\|_2 \leq 1\} \ .$$

Hence, we get by Bousquet's inequality (7.48) with $\mathcal{F} = \{s \ ; \ s \in M, \ \|s\|_2 \leq 1\}$, for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi_M \geq \sqrt{2\sigma^2 \frac{x}{n}} + (1+\delta)\mathbb{E}[\chi_M] + \left(\frac{1}{3} + \frac{1}{\delta}\right)\frac{bx}{n}\right] \leq \exp(-x) \tag{5.121}$$

where

$$\sigma^2 \leq \sup_{s \in M, \ \|s\|_2 \leq 1} \mathbb{E}\left[(s(X))^2\right] = 1$$

and

$$b \leq \sup_{s \in M, \ \|s\|_2 \leq 1} \|s - P(s)\|_\infty \leq 2 \sup_{s \in M, \ \|s\|_2 \leq 1} \|s\|_\infty \leq 2 r_M \sqrt{D_M} \qquad \text{by (5.91).}$$

Moreover, since

$$\sum_{I\in\Lambda_M} \mathrm{Var}\left(\varphi_I\right) = \sum_{I\in\Lambda_M} \left(1 - P\left(I\right)\right) = D_M - 1 \ ,$$

it holds

$$\mathbb{E}\left[\chi_M\right] \leq \sqrt{\mathbb{E}\left[\chi_M^2\right]} = \sqrt{\frac{\sum_{I\in\Lambda_M} \mathrm{Var}\left(\varphi_I\right)}{n}} = \sqrt{\frac{D_M - 1}{n}} \ .$$

So, from (5.121) it follows that

$$\mathbb{P}\left[\chi_M \geq \sqrt{\frac{2x}{n}} + (1+\delta)\sqrt{\frac{D_M-1}{n}} + \left(\frac{1}{3} + \frac{1}{\delta}\right)\frac{2r_M\sqrt{D_M}x}{n}\right] \leq \exp\left(-x\right) \ . \tag{5.122}$$

Hence, taking $x = \beta \ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ in (5.122), we can derive that a positive constant $L_{r_M,\beta}$ exists such that

$$\mathbb{P}\left[\chi_M \geq \left(1 + L_{r_M,\beta}\left(\sqrt{\frac{\ln n}{D_M}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{D_M-1}{n}}\right] \leq n^{-\beta} \ ,$$

which gives the result.■

Lemma 5.9 *Let $r > 1$ and $C > 0$. Assume that (**Abd**) and (**Alr**) hold. If positive constants A_-, A_+, A_l, A_u exist such that*

$$A_+\frac{n}{\left(\ln n\right)^2} \geq D \geq A_-\left(\ln n\right)^2 \quad \text{and} \quad A_l\frac{D}{n} \leq rC \leq A_u\frac{D}{n} \ ,$$

and if the constant A_∞ defined in (5.78) satisfies

$$A_\infty \geq 64 B_2\sqrt{A_u}A_* r_M \ , \tag{5.123}$$

then a positive constant $L_{A_l,A_u,A_{\min}}$ exists such that, for all $n \geq n_0\left(B_2, A_+, A_-, A_l, A_u, r_M, A_{\min}\right)$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right] \geq \left(1 - \frac{L_{A_l,A_u,A_{\min}}}{\sqrt{D_M}}\right)\sqrt{\frac{2rC\left(D_M-1\right)}{n}} \ . \tag{5.124}$$

In the previous Lemma, we state a sharp lower bound for the mean of the supremum of the empirical process on the linear parts of contrasted functions of $\widetilde{M}$ belonging to a slice of excess risk. This is done for models of reasonable dimensions. Moreover, we see that we need to assume that the constant A_∞ introduced in (5.78) is large enough. In order to prove Lemma 5.9 we need the following intermediate result.

Lemma 5.10 *Let $r > 1$, $A_+, A_-, A_u, \beta > 0$ and $C \geq 0$. Assume that (**Abd**) and (**Alr**) hold and that*

$$A_+\frac{n}{\left(\ln n\right)^2} \geq D_M \geq A_-\left(\ln n\right)^2 \quad \text{and} \quad rC \leq A_u\frac{D_M}{n} \ .$$

Set

$$\beta_{n,I} = \frac{\sqrt{2rC}\left(P_n - P\right)\left(\varphi_I\right)}{\sqrt{\sum_{I\in\Lambda_M}\left(P_n - P\right)^2\left(\varphi_I\right)}} \qquad \text{for all } I \in \Lambda_M \ ,$$

and

$$s_{CS} = \sum_{I\in\Lambda_M} \beta_{n,I}\varphi_I \ \in M \ .$$

Then the following inequality holds,

$$\int_{\mathcal{Z}} (s_M s_{CS} + s_M)\, d\mu = 1 \tag{5.125}$$

and if the constant A_∞ defined in (5.78) satisfies

$$A_\infty \geq 32 B_2 \sqrt{A_u \beta} A_* r_M \ ,$$

then it holds, for all $n \geq n_0 (B_2, A_+, A_-, r_M, \beta)$,,

$$\mathbb{P}\left[\max_{I \in \Lambda_M} |\beta_{n,I}| \geq \frac{\tilde{R}_{n, D_M, \alpha}}{A_* r_M \sqrt{D_M}} \right] \leq \frac{2D_M + 1}{n^\beta} \ . \tag{5.126}$$

In this case, $(s_M \times s_{CS} + s_M) \in \mathcal{F}_{(C, rC]}$ with probability at least $1 - (2D_M + 1)\, n^{-\beta}$.

Proof of Lemma 5.10. Let us begin with property (5.125). As $\int_{\mathcal{Z}} s_M d\mu = 1$, it suffices to check that

$$\int_{\mathcal{Z}} s_M \times s_{CS} d\mu = 0 \ .$$

Indeed, as by (5.7) we have $s_M = \sum_{I \in \Lambda_M} \frac{P(I)}{\mu(I)} \mathbf{1}_I$,

$$\begin{aligned} s_M \times s_{CS} &= \frac{\sqrt{2rC}}{\sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I)}} \sum_{I \in \Lambda_M} (P_n - P) \left(\frac{\mathbf{1}_I}{\sqrt{P(I)}} \right) \frac{P(I)}{\mu(I)} \frac{\mathbf{1}_I}{\sqrt{P(I)}} \\ &= \frac{\sqrt{2rC}}{\sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I)}} \sum_{I \in \Lambda_M} (P_n - P)(\mathbf{1}_I) \frac{\mathbf{1}_I}{\mu(I)} \ . \end{aligned}$$

So the expectation of $s_M \times s_{CS}$ with respect to μ is proportional to

$$\begin{aligned} &\int_{\mathcal{Z}} \sum_{I \in \Lambda_M} (P_n - P)(\mathbf{1}_I) \frac{\mathbf{1}_I}{\mu(I)} d\mu \\ &= (P_n - P)(1_{\mathcal{Z}}) = 0 \ . \end{aligned}$$

Thus property (5.125) is satisfied. We now turn to the proof of (5.126). As in the proof of Lemma 5.8 we write

$$\chi_M = \sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I)} \ .$$

By Cauchy-Schwarz inequality, we get

$$\chi = \sup_{s \in S_M} |(P_n - P)(s)| \ ,$$

where S_M is the unit sphere of M, that is

$$S_M = \left\{ s \in M, s = \sum_{I \in \Lambda_M} \beta_I \varphi_I \text{ and } \sqrt{\sum_{I \in \Lambda_M} \beta_I^2} = 1 \right\} \ .$$

Thus we can apply Klein-Rio's bound (7.50) to χ since it holds

$$\begin{aligned} \sup_{s\in S_M} \|s - Ps\|_\infty &\leq 2 \sup_{s\in S_M} \|s\|_\infty \leq 2r_M\sqrt{D_M} \qquad \text{by (5.91).} \\ \sup_{s\in S_M} \mathrm{Var}\left(s\right) &\leq 1 \end{aligned} \tag{5.127}$$

and also, by Inequality (7.45), using (5.127),

$$\begin{aligned} \mathbb{E}\left[\chi_M\right] &\geq B_2^{-1}\sqrt{\mathbb{E}\left[\chi_M^2\right]} - \frac{2r_M\sqrt{D_M}}{n} \\ &= B_2^{-1}\sqrt{\frac{D_M-1}{n}} - \frac{2r_M\sqrt{D_M}}{n} \ . \end{aligned}$$

We thus obtain, for all $\varepsilon, x > 0$,

$$\mathbb{P}\left[\chi_M \leq \left(1-\varepsilon\right) B_2^{-1}\sqrt{\frac{D_M-1}{n}} - \sqrt{2\frac{x}{n}} - \left(1-\varepsilon+\left(1+\frac{1}{\varepsilon}\right)x\right)\frac{2r_M\sqrt{D_M}}{n}\right] \leq \exp\left(-x\right) \ .$$

So, by taking $\varepsilon = \frac{1}{2}$ and $x = \beta\ln n$, and as $D_M \geq A_-\left(\ln n\right)^2$, it holds, for all $n \geq n_0\left(B_2, A_-, r_M, \beta\right)$,

$$\mathbb{P}\left[\chi_M \leq \frac{B_2^{-1}}{8}\sqrt{\frac{D_M}{n}}\right] \leq n^{-\beta} \ . \tag{5.128}$$

Furthermore, combining Bernstein's inequality (7.46) with the observation that we have, for every $I \in \Lambda_M$,

$$\begin{aligned} \|\varphi_I - P\varphi_I\|_\infty &\leq 2\|\varphi_I\|_\infty \leq 2r_M\sqrt{D_M} \qquad \text{by (5.92)} \\ \mathrm{Var}\left(\varphi_I\right) &\leq 1 \ , \end{aligned}$$

we get that, for every $x > 0$,

$$\mathbb{P}\left[\left|\left(P_n - P\right)\left(\varphi_I\right)\right| \geq \sqrt{2\frac{x}{n}} + \frac{2r_M\sqrt{D_M}x}{3n}\right] \leq 2\exp\left(-x\right) \ .$$

Hence, for $x = \beta\ln n$, it comes

$$\mathbb{P}\left[\max_{I\in\Lambda_M}\left|\left(P_n - P\right)\left(\varphi_I\right)\right| \geq \sqrt{\frac{2\beta\ln n}{n}} + \frac{2r_M\sqrt{D_M}\beta\ln n}{3n}\right] \leq \frac{2D_M}{n^\beta} \ , \tag{5.129}$$

then by using (5.128) and (5.129), for all $n \geq n_0\left(B_2, A_-, r_M, \beta\right)$,

$$\mathbb{P}\left[\max_{I\in\Lambda_M}\left|\beta_{n,I}\right| \geq \frac{8B_2\sqrt{2rC}}{\sqrt{\frac{D_M}{n}}}\left(\sqrt{\frac{2\beta\ln n}{n}} + \frac{2r_M\sqrt{D_M}\beta\ln n}{3n}\right)\right] \leq \frac{2D_M+1}{n^\beta} \ .$$

Finally, as $A_+\frac{n}{\left(\ln n\right)^2} \geq D_M$ we have, for all $n \geq n_0\left(A_+, r_M, \beta\right)$,

$$\frac{2r_M\sqrt{D_M}\beta\ln n}{3n} \leq \sqrt{\frac{2\beta\ln n}{n}}$$

and we can check that if

$$A_\infty \geq 32B_2\sqrt{A_u\beta}A_*r_M$$

then, for all $n \geq n_0\left(B_2, A_+, A_-, r_M, \beta\right)$,

$$\mathbb{P}\left[\max_{I \in \Lambda_M} |\beta_{n,I}| \geq \frac{A_\infty}{A_* r_M}\sqrt{\frac{\ln n}{n}}\right] \leq \frac{2D_M+1}{n^\beta} \ .$$

which readily yields Inequality (5.126). As a consequence, it holds with probability at least $1-(2D_M+1)\,n^{-\beta}$,

$$\begin{aligned}
\left\|(s_M \times s_{CS} + s_M) - s_M\right\|_\infty &\leq \|s_M\|_\infty \|s_{CS}\|_\infty \\
&\leq A_* \|s_{CS}\|_\infty && \text{by (5.44) and (5.7)} \\
&= A_* \left\| \sum_{I \in \Lambda_M} \beta_{n,I} \varphi_I \right\|_\infty \\
&\leq A_* r_M \sqrt{D_M} \max_{I \in \Lambda_M} |\beta_{n,I}| && \text{by (5.92)} \\
&\leq \tilde{R}_{n,D_M,\alpha} && \text{by (5.126)} \qquad (5.130)
\end{aligned}$$

Now, by observing that

$$\begin{aligned}
\left\|\psi_{1,M} \cdot \left((s_M \times s_{CS} + s_M) - s_M\right)\right\|_2^2 &= \|s_{CS}\|_2^2 \\
&= 2rC \ ,
\end{aligned}$$

we get by (5.125) and (5.130) that for all $n \geq n_0\left(B_2, A_-, r_M, \beta\right)$, $(s_M \times s_{CS} + s_M) \in \mathcal{F}_{(C,rC]}$ with probability at least $1-(2D_M+1)\,n^{-\beta}$. ■

We are now ready to prove the lower bound (5.124) for the expected value of the largest increment of the empirical process over $\mathcal{F}_{(C,rC]}$.

Proof of Lemma 5.9. Let us begin with the lower bound of $\mathbb{E}^{\frac{1}{2}}\left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right)\right)^2$, a result that will be needed further in the proof. By Lemma 5.10, if we set

$$\tilde{\Omega} = \left\{(s_M \times s_{CS} + s_M) \in \mathcal{F}_{(C,rC]}\right\}$$

if we choose $\beta = 4$ and if

$$A_\infty \geq 64 B_2 \sqrt{A_u} A_* r_M \ ,$$

then it holds, for all $n \geq n_0\left(B_2, A_+, A_-, r_M\right)$,

$$\mathbb{P}\left[\tilde{\Omega}\right] \geq 1 - \frac{2D_M+1}{n^4} \ . \qquad (5.131)$$

Also, it holds

$$\begin{aligned}
&\mathbb{E}^{\frac{1}{2}}\left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right)\right)^2 \\
&\geq \mathbb{E}^{\frac{1}{2}}\left[\left((P_n - P)\left(\sum_{I \in \Lambda_M} \beta_{n,I}\varphi_I\right)\right)^2 \mathbf{1}_{\tilde{\Omega}}\right] \\
&\geq \sqrt{2rC}\sqrt{\mathbb{E}\left[\left(\sum_{I \in \Lambda_M} (P_n - P)^2(\varphi_I)\right)\mathbf{1}_{\tilde{\Omega}}\right]} \ . \qquad (5.132)
\end{aligned}$$

Furthermore, since by (5.92) $\|\varphi_I\|_\infty \leq \sqrt{D_M} r_M$ for all $I \in \Lambda_M$, and since $P(\varphi_I) \geq 0$ we have

$$\left| \sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I) \right| \leq D_M \max_{I \in \Lambda_M} \|\varphi_I\|_\infty^2 \leq r_M^2 D_M^2$$

and it ensures by (5.131), for all $n \geq n_0(B_2, A_+, A_-, r_M)$,

$$\mathbb{E}\left[\left(\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I) \right) \mathbf{1}_{\tilde{\Omega}} \right] \geq \mathbb{E}\left[\left(\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I) \right) \right] - r_M^2 D_M^2 \frac{2D_M + 1}{n^4} .$$

Comparing the last inequality with (5.132), we obtain the lower bound, for all $n \geq n_0(B_2, A_+, A_-, r_M)$,

$$\begin{aligned}
&\mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 \\
&\geq \sqrt{2rC} \sqrt{ \mathbb{E}\left[\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I) \right] } - r_M D_M \sqrt{2rC} \sqrt{\frac{2D_M + 1}{n^4}} \\
&= \sqrt{\frac{2rC(D_M - 1)}{n}} - r_M D_M \sqrt{2rC} \sqrt{\frac{2D_M + 1}{n^4}} .
\end{aligned}$$

Now, since $D_M \leq A_+ n (\ln n)^2$, we get for all $n \geq n_0(A_+, r_M)$,

$$r_M D_M \sqrt{2rC} \sqrt{\frac{2D_M + 1}{n^4}} \leq \frac{1}{\sqrt{D_M}} \times \sqrt{\frac{2rC(D_M - 1)}{n}}$$

and so, if $A_\infty \geq 64 B_2 \sqrt{A_u} A_* r_M$ then, for all $n \geq n_0(B_2, A_+, A_-, r_M)$,

$$\mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 \geq \left(1 - \frac{1}{\sqrt{D_M}} \right) \sqrt{\frac{2rC(D_M - 1)}{n}} . \tag{5.133}$$

Now, as $D_M \geq A_- (\ln n)^2$ we have for all $n \geq n_0(A_-)$, $D_M^{-1/2} \leq 1/2$. Moreover we have $rC \geq A_l D_M n^{-1}$, so we deduce from (5.133) that, for all $n \geq n_0(B_2, A_+, A_-, r_M)$,

$$\mathbb{E}^{\frac{1}{2}} \left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right)^2 \geq \sqrt{\frac{A_l}{2}} \frac{D_M}{n} . \tag{5.134}$$

We turn now to the lower bound of $\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right]$. First observe that $s \in \mathcal{F}_{(C,rC]}$ implies that $2s_M - s \in \mathcal{F}_{(C,rC]}$, so that

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) \right] = \mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} |(P_n - P)(\psi_{1,M} \cdot (s_M - s))| \right] . \tag{5.135}$$

In the next step, we apply Corollary 7.2. More precisely, using notations of Corollary 7.2, we set

$$\mathcal{F} = \left\{ \psi_{1,M} \cdot (s_M - s) , \ s \in \mathcal{F}_{(C,rC]} \right\}$$

and

$$Z = \sup_{s \in \mathcal{F}_{(C,rC]}} |(P_n - P)(\psi_{1,M} \cdot (s_M - s))| .$$

Now, since for all $n \geq n_0(A_+, A_\infty)$, it holds $\tilde{R}_{n,D_M,\alpha} \leq 1/2$, we get by (5.45),

$$\sup_{f\in\mathcal{F}} \|f - Pf\|_\infty \leq 2 \sup_{s\in\mathcal{F}_{(C,rC]}} \|\psi_{1,M}\cdot(s_M - s)\|_\infty \leq 2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha} \leq A_{\min}^{-1} .$$

we set $b = A_{\min}^{-1}$. Since we assume that $rC \leq A_u\frac{D_M}{n}$, it moreover holds

$$\sup_{f\in\mathcal{F}} \mathrm{Var}(f) \leq \sup_{s\in\mathcal{F}_{(C,rC]}} P(\psi_{1,M}\cdot(s_M - s))^2 \leq 2rC \leq 2A_u\frac{D_M}{n}$$

and so we set $\sigma^2 = 2A_u\frac{D_M}{n}$. Now, by (5.134) we have, for all $n \geq n_0(B_2, A_+, A_-, r_M)$,

$$\sqrt{\mathbb{E}[Z^2]} \geq \sqrt{\frac{A_l}{2}}\frac{D_M}{n} . \tag{5.136}$$

Hence, a positive constant $L_{A_l,A_u,A_{\min}}$ exists such that, by setting

$$\varkappa_n = \frac{L_{A_l,A_u,A_{\min}}}{\sqrt{D_M}}$$

we can get using (5.136), that for all $n \geq n_0(B_2, A_+, A_-, r_M)$,

$$\varkappa_n^2\mathbb{E}\left[Z^2\right] \geq \frac{\sigma^2}{n}$$

$$\varkappa_n^2\sqrt{\mathbb{E}[Z^2]} \geq \frac{b}{n}$$

and that, as $D_M \geq A_-(\ln n)^2$, we have for all $n \geq n_0(A_l, A_u, A_-, A_{\min})$,

$$\varkappa_n \in (0,1) .$$

So, using (5.135) and Corollary 7.2, it holds for all $n \geq n_0(B_2, A_+, A_-, A_l, A_u, r_M, A_{\min})$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M}\cdot(s_M - s))\right] \geq \left(1 - \frac{L_{A_l,A_u,A_{\min}}}{\sqrt{D_M}}\right)\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M}\cdot(s_M - s))\right)^2 . \tag{5.137}$$

Finally, using (5.133) in the right-hand side of Inequality (5.137), we can deduce that for all $n \geq n_0(B_2, A_+, A_-, A_l, A_u, r_M, A_{\min})$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M}\cdot(s_M - s))\right] \geq \left(1 - \frac{L_{A_l,A_u,A_{\min}}}{\sqrt{D_M}}\right)\sqrt{\frac{2rC(D_M - 1)}{n}}$$

and so (5.124) is proved. ■

The two following lemmas give some controls of the supremum over the second order terms in the expansion of the contrast (5.81).

Lemma 5.11 *Let $C \geq 0$ and $A_+ > 0$. Under (5.45), assuming that*

$$A_+\frac{n}{(\ln n)^2} \geq D_M ,$$

it holds, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_C} \left|(P_n - P)\left(\psi_2\circ\left(\frac{s - s_M}{s_M}\right)\right)\right|\right] \leq 8A_{\min}^{-2}\tilde{R}_{n,D_M,\alpha}\sqrt{\frac{2C(D_M - 1)}{n}} .$$

Proof. We define the Rademacher process $\mathcal{R}_n$ on a class $\mathcal{F}$ of measurable functions from $\mathcal{Z}$ to $\mathbb{R}$, to be

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(\xi_i) \ , \quad f \in \mathcal{F}$$

where ε_i are independent Rademacher random variables also independent from the ξ_i. By the usual symmetrization argument we have

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s-s_M}{s_M}\right)\right)\right|\right] \leq 2\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2 \circ \left(\frac{s-s_M}{s_M}\right)\right)\right|\right] \ . \tag{5.138}$$

As $A_+ \frac{n}{(\ln n)^2} \geq D_M$, we have, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\tilde{R}_{n,D_M,\alpha} \leq \frac{A_{\min}}{2} \ .$$

Hence, by Inequality (5.19) of Lemma 5.3 it holds for all $n \geq n_0(A_+, A_{\min}, A_\infty)$, for all $(x,y) \in \left[-A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha},\ A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right]^2$,

$$|\psi_2(x) - \psi_2(y)| \leq 2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}|x-y| \ . \tag{5.139}$$

We define now the following real-valued function ρ,

$$\rho(x) = \begin{cases} \left(2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right)^{-1}\psi_2(x) & \text{if } x \in \left[-A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha},\ A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right] \\ \left(2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right)^{-1}\psi_2\left(-A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right) & \text{if } x \leq -A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha} \\ \left(2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right)^{-1}\psi_2\left(A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\right) & \text{if } x \geq A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha} \end{cases}$$

and since $\rho(0) = h(0) = 0$, it follows from (5.139) that ρ is a contraction mapping for all $n \geq n_0(A_+, A_{\min}, A_\infty)$. Then, taking the expectation with respect to the Rademacher variables, we then get for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\begin{aligned} &\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2 \circ \left(\frac{s-s_M}{s_M}\right)\right)\right|\right] \\ &= 2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\rho\left(\left(\frac{s-s_M}{s_M}\right)(\xi_i)\right)\right|\right] \end{aligned} \tag{5.140}$$

We can now apply Theorem 7.4 to get for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\begin{aligned} \mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\rho\left(\left(\frac{s-s_M}{s_M}\right)(\xi_i)\right)\right|\right] &\leq 2\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(\frac{s-s_M}{s_M}\right)(\xi_i)\right|\right] \\ &= 2\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\frac{s-s_M}{s_M}\right)\right|\right] \end{aligned} \tag{5.141}$$

and so we derive successively the following upper bounds in mean, for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right|\right] = \mathbb{E}\left[\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right|\right]\right] \\
&\leq 2A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\mathbb{E}\left[\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\rho\left(\left(\frac{s-s_M}{s_M}\right)(\xi_i)\right)\right|\right]\right] \qquad \text{by (3.187)} \\
&\leq 4A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\frac{s-s_M}{s_M}\right)\right|\right] \qquad \text{by (3.188)} \\
&\leq 4A_{\min}^{-1}\tilde{R}_{n,D_M,\alpha}\sqrt{\mathbb{E}\left[\left(\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\frac{s-s_M}{s_M}\right)\right|\right)^2\right]}\,. \qquad (5.142)
\end{aligned}$$

Hence, for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$,

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[\left(\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\frac{s-s_M}{s_M}\right)\right|\right)^2\right]} \\
&= \sqrt{\mathbb{E}\left[\left(\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_{1,M}\cdot(s_M-s)\right)\right|\right)^2\right]} \\
&\leq \sqrt{\mathbb{E}\left[\left(\sup\left\{\left|\sum_{I\in\Lambda_M}a_I\mathcal{R}_n\left(\varphi_I\right)\right|;\ \sum_{I\in\Lambda_M}a_I^2\leq 2C\right\}\right)^2\right]} \\
&= \sqrt{2C}\sqrt{\mathbb{E}\left[\sum_{I\in\Lambda_M}\left(\mathcal{R}_n\left(\varphi_I\right)\right)^2\right]} = \sqrt{\frac{2C\left(D_M-1\right)}{n}} \qquad (5.143)
\end{aligned}$$

and the result follows by injecting (5.142) and (5.143) in (5.138). ■

Lemma 5.12 *Let $A_+, A_-, A_l, \beta, C_- > 0$, and assume (5.45) and (**Alr**). Then if $C_- \geq A_l\frac{D_M}{n}$ and $A_+ n\left(\ln n\right)^{-2} \geq D_M \geq A_-\left(\ln n\right)^2$, then a positive constant $L_{A_-,A_l,\beta}$ exists such that, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+, A_l\right)$,*

$$\mathbb{P}\left[\forall C > C_-,\quad \sup_{s\in\mathcal{F}_C}\left|\left(P_n-P\right)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| \leq L_{A_-,A_l,A_{\min},\beta}\sqrt{\frac{C\left(D_M-1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \geq 1-n^{-\beta}\,.$$

Proof. First notice that, as $A_+ n\left(\ln n\right)^{-2} \geq D_M$,

$$\tilde{R}_{n,D_M,\alpha} \leq \frac{A_\infty\sqrt{A_+}}{\sqrt{\ln n}}\,.$$

As a consequence, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$,

$$\tilde{R}_{n,D_M,\alpha} \leq \sqrt{2}A_{\min}\,. \qquad (5.144)$$

Now, since $\cup_{C>C_-}\mathcal{F}_C \subset B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right)$ where

$$B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D_M,\alpha}\right) = \left\{s\in M, \left\|s-s_M\right\|_\infty \leq \tilde{R}_{n,D_M,\alpha}\right\},$$

we have by (5.144) and (5.45), for all $s \in \cup_{C>C_-}\mathcal{F}_C$ and for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$,

$$\begin{aligned}
&\frac{1}{2}\left\|\psi_{1,M}\cdot(s-s_M)\right\|_2^2 \\
&\leq \frac{A_{\min}^{-2}}{2}\left\|s-s_M\right\|_\infty^2 \\
&\leq \frac{A_{\min}^{-2}}{2}\tilde{R}_{n,D_M,\alpha}^2 \leq 1 \ .
\end{aligned}$$

We thus have, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$,

$$\cup_{C>C_-}\mathcal{F}_C = \cup_{C_-\wedge 1<C\leq 1}\mathcal{F}_C$$

and by monotonicity of the collection $\mathcal{F}_C$, for some $q>1$ and $J = \left\lfloor \frac{|\ln(C_-\wedge 1)|}{\ln q}\right\rfloor + 1$, it holds

$$\cup_{C_-\wedge 1<C\leq 1}\mathcal{F}_C \subset \cup_{j=0}^{J}\mathcal{F}_{q^jC_-} \ .$$

Simple computations show that, since $D_M \geq 1$ and $C_- \geq A_l\frac{D_M}{n} \geq \frac{A_l}{n}$, one can find a constant $L_{A_l,q}$ such that

$$J \leq L_{A_l,q}\ln n \ .$$

Moreover, by monotonicity of $C \longmapsto \sup_{s\in\mathcal{F}_C}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right|$, we have uniformly in $C \in \left(q^{j-1}C_-, q^jC_-\right]$,

$$\sup_{s\in\mathcal{F}_C}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| \leq \sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| \ .$$

Hence we get, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$ and any $L>0$,

$$\begin{aligned}
&\mathbb{P}\left[\forall C > C_-,\quad \sup_{s\in\mathcal{F}_C}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| \leq L\sqrt{\frac{2C\left(D_M-1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \\
&\geq \mathbb{P}\left[\forall j\in\{1,...,J\},\quad \sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| \leq L\sqrt{\frac{2q^jC_-\left(D_M-1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \ .
\end{aligned}$$

Now, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$ and any $L>0$,

$$\begin{aligned}
&\mathbb{P}\left[\forall j\in\{1,...,J\},\quad \sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| \leq L\sqrt{\frac{2q^jC_-\left(D_M-1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \\
&= 1-\mathbb{P}\left[\exists j\in\{1,...,J\},\quad \sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| > L\sqrt{\frac{2q^jC_-\left(D_M-1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \\
&\geq 1-\sum_{j=1}^{J}\mathbb{P}\left[\sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| > L\sqrt{\frac{2q^jC_-\left(D_M-1\right)}{n}}\tilde{R}_{n,D_M,\alpha}\right] .
\end{aligned} \tag{5.145}$$

Given $j\in\{1,...,J\}$, Lemma 3.10 yields

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{q^jC_-}}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right|\right] \leq 8A_{\min}^{-2}\sqrt{\frac{2q^jC_-\left(D_M-1\right)}{n}}\tilde{R}_{n,D_M,\alpha} \ ,$$

and we can next apply Bousquet's inequality (7.48) to handle the deviations around the mean. Since for all $n \geq n_0(A_{\min}, A_\infty, A_+)$ we have for all $s \in \mathcal{F}_{q^j C_-}$,

$$\|s - s_M\|_\infty \leq \tilde{R}_{n,D_M,\alpha} \leq \frac{A_{\min}}{2}$$

we can apply Inequalities (5.45) and (5.83) to get, for all $n \geq n_0(A_{\min}, A_\infty, A_+)$,

$$\begin{aligned}
\sup_{s \in \mathcal{F}_{q^j C_-}} \left\| \psi_2 \circ \left(\frac{s - s_M}{s_M} \right) - P\left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right\| &\leq 2 \sup_{s \in \mathcal{F}_{q^j C_-}} \|\psi_2^s \cdot (s - s_M)\|_\infty \\
&\leq 2A_{\min}^{-1} \sup_{s \in \mathcal{F}_{q^j C_-}} \left\| \frac{1}{s_M} (s - s_M)^2 \right\|_\infty \\
&\leq 2A_{\min}^{-2} \tilde{R}_{n,D_M,\alpha}^2
\end{aligned}$$

and, for all $s \in \mathcal{F}_{q^j C_-}$, for all $n \geq n_0(A_{\min}, A_\infty, A_+)$,

$$\begin{aligned}
&\mathrm{Var}\left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \\
&\leq P\left[\left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right)^2 \right] \\
&\leq A_{\min}^{-2} \|s - s_M\|_\infty^2 P\left[\left(\frac{s - s_M}{s_M} \right)^2 \right] \qquad \text{by (5.83)} \\
&\leq 2A_{\min}^{-2} \tilde{R}_{n,D_M,\alpha}^2 q^j C_- \ .
\end{aligned}$$

It follows that Inequality (7.48) applied with $\varepsilon = 1$ gives, for all $x > 0$ and for all $n \geq n_0(A_{\min}, A_\infty, A_+)$,

$$\mathbb{P}\left[\begin{array}{c} \sup_{s \in \mathcal{F}_{q^j C_-}} \left| (P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| \geq \\ 16A_{\min}^{-2} \sqrt{\frac{2q^j C_- (D_M - 1)}{n}} \tilde{R}_{n,D_M,\alpha} + \sqrt{\frac{4A_{\min}^{-2} \tilde{R}_{n,D_M,\alpha}^2 q^j C_- x}{n}} + \frac{8A_{\min}^{-2} \tilde{R}_{n,D_M,\alpha}^2 x}{3n} \end{array} \right] \leq \exp(-x) \ . \tag{5.146}$$

As a consequence, as $D_M \geq A_- (\ln n)^2$, $C_- \geq A_l D_M n^{-1}$ and as $\tilde{R}_{n,D_M,\alpha} \leq 1$ for all $n \geq n_0(A_{\min}, A_\infty, A_+)$, taking $x = \gamma \ln n$ in (5.146) for some $\gamma > 0$, easy computations show that a positive constant $L_{A_-, A_l, A_{\min}, \gamma}$ independent of j exists such that for all $n \geq n_0(A_{\min}, A_\infty, A_+)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{q^j C_-}} \left| (P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| \geq L_{A_-, A_l, A_{\min}, \gamma} \sqrt{\frac{q^j C_- (D_M - 1)}{n}} \tilde{R}_{n,D_M,\alpha} \right] \leq \frac{1}{n^\gamma} \ .$$

Hence, using (3.190), we get for all $n \geq n_0(A_{\min}, A_\infty, A_+)$,

$$\mathbb{P}\left[\forall C > C_-, \quad \sup_{s \in \mathcal{F}_C} \left| (P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| \leq L_{A_-, A_l, A_{\min}, \gamma} \sqrt{\frac{2C (D_M - 1)}{n}} \tilde{R}_{n,D_M,\alpha} \right] \geq 1 - \frac{J}{n^\gamma} \ .$$

And finally, as $J \leq L_{A_l,q} \ln n$, taking $\gamma = \beta + 1$ and $q = 2$ gives the result for all $n \geq n_0(A_{\min}, A_\infty, A_+, A_l)$.

■

Having controlled the residual empirical process driven by the remainder terms in the contrast, and having proved sharp bounds for the expectation of the increments of the main empirical process on our slices, it remains to combine the above lemmas in order to establish the crucial probability estimates controlling the empirical excess risk on the slides.

Lemma 5.13 *Let $\beta, A_-, A_+, A_l, C > 0$. Assume that (5.45) and (**Alr**) hold. A positive constant A_4 exists, only depending on $A_{\min}, A_\infty, r_M, \beta$, such that, if*

$$A_l \frac{D_M}{n} \leq C \leq (1 + A_4 \nu_n)^2 \frac{D_M - 1}{2n} \quad \text{and} \quad A_+ \frac{n}{(\ln n)^2} \geq D_M \geq A_- (\ln n)^2$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D_M}},\ \sqrt{\frac{D_M \ln n}{n}}\right\}$, then for all $n \geq n_0 (A_l, A_-, A_+, A_{\min}, r_M, A_\infty, \beta)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_C} P_n (K s_M - K s) \geq \left(1 + L_{A_-, A_l, A_{\min}, A_\infty, r_M, \beta} \times \nu_n\right) \sqrt{\frac{2C (D_M - 1)}{n}} - C\right] \leq 2 n^{-\beta} .$$

Proof. Start with

$$\begin{aligned}
&\sup_{s \in \mathcal{F}_C} P_n (K s_M - K s) \\
&= \sup_{s \in \mathcal{F}_C} \left\{ P_n \left(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right\} \\
&= \sup_{s \in \mathcal{F}_C} \left\{ (P_n - P) (\psi_{1,M} \cdot (s_M - s)) - (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) - P (K s - K s_M) \right\} \\
&\leq \sup_{s \in \mathcal{F}_C} \{ (P_n - P) (\psi_{1,M} \cdot (s_M - s)) - P (K s - K s_M) \} \\
&+ \sup_{s \in \mathcal{F}_C} \left| (P_n - P) \left(\psi_2 \circ \left(\frac{s - s_M}{s_M} \right) \right) \right| . \qquad (5.147)
\end{aligned}$$

Recall that by (5.89) we have, for all $s \in \mathcal{F}_C$ and for all $n \geq n_0 (A_+, A_{\min}, A_\infty)$,

$$P (K s - K s_M) \geq \left(1 - \frac{4}{3 A_{\min}} \tilde{R}_{n, D_M, \alpha}\right) \frac{1}{2} \left\| \psi_{1,M} (s - s_M) \right\|_2^2 . \qquad (5.148)$$

Next, recall that

$$D_L = \left\{ s \in \widetilde{M} \ ;\ \frac{1}{2} \left\| \psi_{1,M} \cdot (s - s_M) \right\|_2^2 = L \right\} \bigcap B_{(M, L_\infty)} \left(s_M, \tilde{R}_{n, D_M, \alpha} \right) .$$

Moreover, we notice that, for any $s \in M$,

$$\psi_{1,M} (s - s_M) = \frac{s - s_M}{s_M}$$

is a piecewise constant function with respect to the partition Λ_M. Thus $\psi_{1,M} \cdot (s - s_M) \in M$ for any $s \in M$, and we have

$$\begin{aligned}
&\sup_{s \in D_L} (P_n - P) (\psi_{1,M} \cdot (s_M - s)) \\
&\leq \sup_{\left\{ s \in M, \|t\|_2^2 = 2L \right\}} (P_n - P) (t) \\
&\leq \sqrt{2L} \sqrt{\sum_{I \in \Lambda_M} (P_n - P)^2 (\varphi_I)}
\end{aligned}$$

where the last bound follows from Cauchy-Schwarz inequality. Then, for all $n \geq n_0(A_+, A_{\min}, A_\infty)$,

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_C}\left\{(P_n-P)(\psi_{1,M}\cdot(s_M-s))-P(Ks-Ks_M)\right\}\\
&\leq \sup_{L\leq C}\sup_{s\in D_L}\left\{(P_n-P)(\psi_{1,M}\cdot(s_M-s))-\left(1-\frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)L\right\} \qquad \text{by (5.148)}\\
&\leq \sup_{L\leq C}\left\{\sqrt{2L}\sqrt{\sum_{I\in\Lambda_M}(P_n-P)^2(\varphi_I)}-\left(1-\frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)L\right\}\ .
\end{aligned}$$

Hence, since $D_M \geq A_-(\ln n)^2 \geq 2$ for all $n \geq n_0(A_-)$, we deduce from Lemma 5.8 that for all $n \geq n_0(A_-, A_+, A_{\min}, A_\infty)$,

$$\mathbb{P}\left[\begin{array}{c}\sup_{s\in\mathcal{F}_C}\left\{(P_n-P)(\psi_{1,M}\cdot(s_M-s))-P(Ks-Ks_M)\right\}\\ \geq \sup_{L\leq C}\left\{\sqrt{2L}(1+\tau_n)\sqrt{\frac{D_M-1}{n}}-\left(1-\frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\right)L\right\}\end{array}\right] \leq n^{-\beta}\ . \tag{5.149}$$

where

$$\begin{aligned}
\tau_n &= L_{r_M,\beta}\left(\sqrt{\frac{\ln n}{D_M}}\vee\frac{\sqrt{\ln n}}{n^{1/4}}\right)\\
&\leq L_{r_M,\beta}\left(\sqrt{\frac{\ln n}{D_M}}\vee\sqrt{\frac{D_M\ln n}{n}}\right)\\
&\leq L_{r_M,\beta}\nu_n\ .
\end{aligned} \tag{5.150}$$

Assume now that

$$C \leq \frac{D_M-1}{n}\ . \tag{5.151}$$

then we have for all $0 \leq L \leq C$,

$$\frac{4}{3A_{\min}}\tilde{R}_{n,D_M,\alpha}\times L \leq L_{A_{\min},A_\infty}\sqrt{\frac{D_M\ln n}{n}}\times\sqrt{L}\sqrt{\frac{D_M-1}{n}} \leq L_{A_{\min},A_\infty}\nu_n\sqrt{L}\sqrt{\frac{D_M-1}{n}}\ . \tag{5.152}$$

Hence, using (5.150) and (5.152) in (5.149), if $C \leq \frac{D_M-1}{n}$ it holds for all $n \geq n_0(A_-, A_+, A_{\min}, A_\infty)$,

$$\mathbb{P}\left[\begin{array}{c}\sup_{s\in\mathcal{F}_C}\left\{(P_n-P)(\psi_{1,M}\cdot(s_M-s))-P(Ks-Ks_M)\right\}\\ \geq \sup_{L\leq C}\left\{\sqrt{2L}(1+L_{A_{\min},A_\infty,r_M,\beta}\nu_n)\sqrt{\frac{D_M-1}{n}}-L\right\}\end{array}\right] \leq n^{-\beta}\ . \tag{5.153}$$

Now, we set $A_4 = L_{A_{\min},A_\infty,r_M,\beta}$ the positive constant appearing in (5.153). If $C \leq (1+A_4\nu_n)^2\frac{D_M-1}{2n}$ then for all $n \geq n_0(A_4)$ (5.151) is satisfied and we get after simple calculations that

$$\sup_{L\leq C}\left\{\sqrt{2L}(1+A_4\nu_n)\sqrt{\frac{D_M-1}{n}}-L\right\} = \sqrt{2C}(1+A_4\nu_n)\sqrt{\frac{D_M-1}{n}}-C$$

and as a consequence, for all $n \geq n_0(A_-, A_+, A_{\min}, A_4, A_\infty)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left\{(P_n-P)(\psi_{1,M}\cdot(s_M-s))-P(Ks-Ks_M)\right\} \geq \sqrt{2C}(1+A_4\nu_n)\sqrt{\frac{D_M-1}{n}}-C\right] \leq n^{-\beta}\ . \tag{5.154}$$

Moreover, since $C \geq A_l\frac{D_M}{n}$, we can derive from Lemma 5.12 that for all $n \geq n_0(A_{\min}, A_\infty, A_+, A_l)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left|(P_n-P)\left(\psi_2\circ\left(\frac{s-s_M}{s_M}\right)\right)\right| \geq L_{A_-,A_l,A_{\min},\gamma}\sqrt{\frac{C(D_M-1)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \leq n^{-\beta}$$

and as

$$\tilde{R}_{n,D_M,\alpha} \leq L_{A_\infty} \nu_n$$

we have, for all $n \geq n_0 (A_{\min}, A_\infty, A_+, A_l)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C} \left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \geq L_{A_-,A_l,A_{\min},A_\infty} \sqrt{\frac{C(D_M - 1)}{n}} \tilde{R}_{n,D_M,\alpha}\right] \leq n^{-\beta} . \tag{5.155}$$

The conclusion follows by making use of (3.200) and (5.155) in Inequality (5.147). ■

Lemma 5.14 *Let $\beta, A_-, A_+, A_u, C \geq 0$. Assume that (5.45) and (**Alr**) hold. A positive constant A_5, depending on $A_\infty, r_M, A_{\min}, A_-, A_u, \beta$, exists such that, if it holds*

$$A_u \frac{D_M}{n} \geq C \geq \frac{1}{4}(1 + A_5\nu_n)^2 \frac{D_M - 1}{n} \quad \text{and} \quad A_+ \frac{n}{(\ln n)^2} \geq D_M \geq A_- (\ln n)^2$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D_M}}, \sqrt{\frac{D_M \ln n}{n}}\right\}$, then for all $n \geq n_0 (A_\infty, A_{cons}, n_1, A_+, \alpha)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n (Ks_M - Ks) \geq (1 + A_5\nu_n)\sqrt{\frac{2C(D_M - 1)}{n}} - C\right] \leq 2n^{-\beta} .$$

Moreover, when we only assume $C \geq 0$ (and keep the other assumptions unchanged), a positive constant A_6 exists, depending only on $A_\infty, r_M, A_{\min}, A_-, \beta$, such that we have for all $n \geq n_0\left(A_{\min}, A_\infty, A_+, A_-, \tilde{A}_5\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n (Ks_M - Ks) \geq (1 + A_5\nu_n)^2 \frac{D_M - 1}{2n}\right] \leq 2n^{-\beta} . \tag{5.156}$$

Proof. The proof is similar to that of Lemma 5.13 and follows from the same kind of computations. First observe that

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_{>C}} P_n (Ks_M - Ks) \\
&= \sup_{s\in\mathcal{F}_{>C}} \left\{P_n\left(\psi_{1,M} \cdot (s_M - s) - \psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right\} \\
&= \sup_{s\in\mathcal{F}_{>C}} \left\{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right) - P(Ks - Ks_M)\right\} \\
&= \sup_{s\in\mathcal{F}_{>C}} \left\{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - P(Ks - Ks_M) - (P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right\} \\
&\leq \sup_{L>C} \sup_{s\in D_L} \left\{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (1 - L_{A_{\min},A_\infty}\nu_n) L - (P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right\} \text{ by (5.89)} \\
&\leq \sup_{L>C} \left\{\sqrt{2L}\sqrt{\sum_{I\in\Lambda_M} (P_n - P)^2(\varphi_I)} - (1 - L_{A_{\min},A_\infty}\nu_n) L + \sup_{s\in\mathcal{F}_L} \left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right|\right\}
\end{aligned} \tag{5.157}$$

where the last bound follows from Cauchy-Schwarz inequality. From Lemma 5.8 and since for all $n \geq n_0 (A_-)$, $D_M \geq A_- (\ln n)^2 \geq 2$, we can deduce that for all $n \geq n_0 (A_-)$,

$$\mathbb{P}\left[\sqrt{\sum_{I\in\Lambda_M} (P_n - P)^2(\varphi_I)} \geq (1 + L_{r_M,\beta}\nu_n)\sqrt{\frac{D_M - 1}{n}}\right] \leq n^{-\beta} . \tag{5.158}$$

Now, since

$$C \geq \frac{D_M}{2n}$$

we can apply Lemma 5.12 with $A_l = 1/2$, and deduce that, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+\right)$,

$$\mathbb{P}\left[\forall L > C, \quad \sup_{s\in\mathcal{F}_L}\left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \geq L_{A_\infty, A_{\min}, A_-, \beta} \times \nu_n \sqrt{\frac{L\left(D_M - 1\right)}{n}}\right] \leq n^{-\beta} \tag{5.159}$$

Now using (5.158) and (5.159) in (5.157) we obtain, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+, A_-\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \sup_{L>C}\left\{\left(1 + L_{A_\infty, r_M, A_{\min}, A_-, \beta} \times \nu_n\right)\sqrt{\frac{2L\left(D_M - 1\right)}{n}} - \left(1 - L_{r_M,\beta}\right)L\right\}\right] \leq 2n^{-\beta} \tag{5.160}$$

and we set $\tilde{A}_5 = L_{A_\infty, r_M, A_{\min}, A_-, \beta} \vee L_{r_M,\beta}$ where $L_{A_\infty, r_M, A_{\min}, A_-, \beta}$ and $L_{r_M,\beta}$ are the constants appearing in (5.160). Since, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_5\right)$, $0 < \frac{1+\tilde{A}_5\nu_n}{1-\tilde{A}_5\nu_n} \leq 1 + 4\tilde{A}_5\nu_n$, and for $C \geq \left(1 + 4\tilde{A}_5\nu_n\right)^2 \frac{D_M - 1}{2n}$ we get by simple calculations, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_5\right)$,

$$\sup_{L>C}\left\{\sqrt{2L}\left(1 + \tilde{A}_5\nu_n\right)\sqrt{\frac{D_M - 1}{n}} - \left(1 - \tilde{A}_5\nu_n\right)L\right\} = \left(1 + \tilde{A}_5\nu_n\right)\sqrt{\frac{2C\left(D_M - 1\right)}{n}} - \left(1 - \tilde{A}_5\nu_n\right)C \ .$$

Moreover, we have $C \leq A_u\frac{D_M}{n}$, so for all $n \geq n_0\left(A_-\right)$, $C \leq \sqrt{\frac{2A_uC(D_M-1)}{n}}$ and and as a consequence, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_5\right)$,

$$\sup_{L>C}\left\{\sqrt{2L}\left(1 + \tilde{A}_5\nu_n\right)\sqrt{\frac{D_M - 1}{n}} - \left(1 - \tilde{A}_5\nu_n\right)L\right\} \leq \left(1 + \left(1 + \sqrt{A_u}\right)\tilde{A}_5\nu_n\right)\sqrt{\frac{2C\left(D_M - 1\right)}{n}} - C \ ,$$

so, for all $n \geq n_0\left(A_{\min}, A_\infty, A_+, A_-, \tilde{A}_5\right)$,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \left(1 + \left(1 + \sqrt{A_u}\right)\tilde{A}_5\nu_n\right)\sqrt{\frac{2C\left(D_M - 1\right)}{n}} - C\right] \leq 2n^{-\beta}$$

which gives the first part of the lemma by setting $A_5 = 4\tilde{A}_5 \vee \left(1 + \sqrt{A_u}\right)\tilde{A}_5$. The second part comes from (5.160) and the fact that, for any value of $C \geq 0$, for all $n \geq n_0\left(A_+, A_-, \tilde{A}_5\right)$,

$$\sup_{L>C}\left\{\sqrt{2L}\left(1 + \tilde{A}_5\nu_n\right)\sqrt{\frac{D_M - 1}{n}} - \left(1 - \tilde{A}_5\nu_n\right)L\right\} \leq \left(1 + 4\tilde{A}_5\nu_n\right)^2\frac{D_M - 1}{2n} \ .$$

■

Lemma 5.15 *Let $r > 1$ and $C, \beta > 0$. Assume that (**Abd**) and (**Alr**) hold. If positive constants A_-, A_+, A_l, A_u exist such that*

$$A_+\frac{n}{\left(\ln n\right)^2} \geq D_M \geq A_-\left(\ln n\right)^2 \quad \text{and} \quad A_l\frac{D_M}{n} \leq rC \leq A_u\frac{D_M}{n} \ ,$$

and if the constant A_∞ defined in (5.78) satisfies

$$A_\infty \geq 64B_2\sqrt{A_u}A_*r_M \ ,$$

then a positive constant $L_{A_-,A_l,A_u,A_{\min},A_\infty,\beta}$ *exists such that, for all* $n \geq n_0\left(A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty\right)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_-,A_l,A_u,A_{\min},A_\infty,\beta} \times \nu_n\right)\sqrt{\frac{2rC\left(D_M-1\right)}{n}} - rC\right) \leq 2n^{-\beta}\ ,$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D_M}},\ \sqrt{\frac{D_M \ln n}{n}}\right\}$.

Proof. Start with

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \\
&= \sup_{s\in\mathcal{F}_{(C,rC]}} \left\{\left(P_n - P\right)\left(Ks_M - Ks\right) + P\left(Ks_M - Ks\right)\right\} \\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_2 \circ \left(\frac{s-s_M}{s_M}\right)\right) - \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(Ks - Ks_M\right) \\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - \sup_{s\in\mathcal{F}_{rC}} \left(P_n - P\right)\left(\psi_2 \circ \left(\frac{s-s_M}{s_M}\right)\right) - rC \qquad (5.161)
\end{aligned}$$

and set

$$\begin{aligned}
S_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) \\
M_{1,r,C} &= \mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right] \\
b_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|\psi_{1,M}\cdot\left(s_M - s\right) - P\psi_{1,M}\cdot\left(s_M - s\right)\right\|_\infty \\
\sigma^2_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \mathrm{Var}\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\ .
\end{aligned}$$

By Klein-Rio's Inequality (7.50), we get, for all $\delta, x > 0$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1-\delta\right)M_{1,r,C} - \sqrt{\frac{2\sigma^2_{1,r,C}x}{n}} - \left(1+\frac{1}{\delta}\right)\frac{b_{1,r,C}x}{n}\right) \leq \exp\left(-x\right). \qquad (5.162)$$

Then, notice that all conditions of Lemma 5.9 are satisfied and that it gives for all $n \geq n_0\left(A_+, A_-, A_l, A_u, r_M, A_{\min}\right)$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right] \geq \left(1 - \frac{L_{A_l,A_u,A_{\min}}}{\sqrt{D_M}}\right)\sqrt{\frac{2rC\left(D_M-1\right)}{n}}\ . \qquad (5.163)$$

In addition, observe that

$$\sigma^2_{1,r,C} \leq \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(\psi^2_{1,M}\left(s_M - s\right)^2\right) \leq 2rC \qquad (5.164)$$

and for all $n \geq n_0\left(A_+, A_{\min}, A_\infty\right)$,

$$b_{1,r,C} \leq 2\sup_{s\in\mathcal{F}_{(C,rC]}} \left\|\psi_{1,M}\cdot\left(s_M - s\right)\right\|_\infty \leq 2A^{-1}_{\min}\tilde{R}_{n,D_M,\alpha} \leq 1 \qquad (5.165)$$

Hence, using (5.163), (5.164) and (5.165) in Inequality (5.162), we get for all $x > 0$ and all $n \geq n_0 (A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty)$,

$$\mathbb{P}\left(S_{1,r,C} \leq (1-\delta)\left(1 - \frac{L_{A_l,A_u,A_{\min}}}{\sqrt{D_M}}\right)\sqrt{\frac{2rC\,(D_M-1)}{n}} - \sqrt{\frac{4rCx}{n}} - \left(1+\frac{1}{\delta}\right)\frac{x}{n}\right) \leq \exp(-x) \ .$$

Now, taking $x = \beta \ln n$, $\delta = \sqrt{\frac{\ln n}{D_M}}$, we can deduce by simple computations that a positive constant $L_{A_l,A_u,A_{\min},\beta}$ exists such that, for all $n \geq n_0 (A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty)$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1 - L_{A_l,A_u,A_{\min},\beta}\sqrt{\frac{\ln n}{D_M}}\right)\sqrt{\frac{2rC\,(D_M-1)}{n}}\right) \leq n^{-\beta} \tag{5.166}$$

and as

$$\sqrt{\frac{\ln n}{D_M}} \leq \nu_n \ ,$$

(5.166) gives, for all $n \geq n_0 (A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty)$,

$$\mathbb{P}\left(S_{1,r,C} \leq (1 - L_{A_l,A_u,A_{\min},\beta}\nu_n)\sqrt{\frac{2rC\,(D_M-1)}{n}}\right) \leq n^{-\beta} \ . \tag{5.167}$$

Moreover, from Lemma 5.12 we can deduce that, for all $n \geq n_0 (A_{\min}, A_\infty, A_+, A_l)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{rC}} \left| (P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \geq L_{A_-,A_l,A_{\min},\beta}\sqrt{\frac{rC\,(D_M-1)}{n}}\tilde{R}_{n,D_M,\alpha}\right] \leq n^{-\beta} \tag{5.168}$$

and noticing that

$$\tilde{R}_{n,D,\alpha} = A_\infty\sqrt{\frac{D\ln n}{n}} \leq A_\infty \nu_n$$

we deduce from (3.210) that, for all $n \geq n_0 (A_{\min}, A_\infty, A_+, A_l)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{rC}} \left| (P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right| \geq L_{A_-,A_l,A_{\min},A_\infty,\beta}\nu_n\sqrt{\frac{2rC\,(D_M-1)}{n}}\right] \leq n^{-\beta} \ . \tag{5.169}$$

Finally, using (5.167) and (5.169) in (5.161) we get that, for all $n \geq n_0 (A_+, A_-, A_l, A_u, r_M, A_{\min}, A_\infty)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{(C,rC]}} P_n (Ks_M - Ks) \leq \left(1 - L_{A_-,A_l,A_u,A_{\min},A_\infty,\beta} \times \nu_n\right)\sqrt{\frac{2rC\,(D_M-1)}{n}} - rC\right) \leq 2n^{-\beta},$$

which concludes the proof. ■

Chapitre 6

Excess Risks bounds for least-squares estimation of density

This chapter is devoted to least-squares estimation of density. For a detailed introduction to penalized least-squares estimation of density from a nonasymptotic point of view we refer to Chapter 7 of [61], and especially to Section 7.2.
In some recent works, Lerasle ([55], [56], [57]) studied the efficiency of some penalized least-squares estimation procedures of density. The author validates in [56] the slope heuristics first formulated by Birgé and Massart [23] in a generalized Gaussian linear model setting and also proves nonasymptotic quasi-optimality of resampling penalties proposed by Arlot [5] in a regression framework. Indeed, Lerasle shows for the latter penalties pathwise oracle inequalities with leading constant almost one. The author also extends these results in [55] for stationary data under various mixing conditions.
For a probability measure of reference μ, we denote by f the density with respect to μ, to be estimated, $\|\cdot\|$ the natural quadratic norm of $L_2(\mu)$ and we set $M \subset L_2(\mu)$ a linear model of finite dimension D. Then for $(\varphi_k)_{k=1}^D$ an orthonormal basis of $(M, \|\cdot\|)$, the linear projection s_M of f onto M can be written

$$s_M = \sum_{k=1}^{D} P(\varphi_k)\,\varphi_k \tag{6.1}$$

and the least-squares estimator s_n on M satisfies

$$s_n = \sum_{k=1}^{D} P_n(\varphi_k)\,\varphi_k \ . \tag{6.2}$$

It thus can be easily derived that the excess risk of the least-squares estimator on M satisfies

$$\begin{aligned} \|s_n - s_M\|^2 &= \sum_{k=1}^{D} (P - P_n)^2(\varphi_k) \\ &= \sup\left\{ (P - P_n)^2(s) : s \in M, \|s\| \leq 1 \right\} \ . \end{aligned} \tag{6.3}$$

Hence, the excess risk of the least-squares estimator on M is equal to the square of the supremum of the empirical process on the unit ball of $(M, \|\cdot\|)$. Moreover, if we set K the least-squares contrast given by (6.6) below, we see by simple computations that

$$\|s_n - s_M\|^2 = P(Ks_n - Ks_M) = P_n(Ks_M - Ks_n) \ , \tag{6.4}$$

so the true excess risk on M is equal to the empirical one. Based on these observations, the theoretical validation of the slope heuristics given in Theorems 2.2 and 2.3 of [56] heavily

relies on sharp deviations bounds for the excess risk of the least-squares estimators. To do so, the author gives a concentration inequality for the square of the empirical process by using Bousquet and Klein inequalities, see Corollary 6.5 of [56].
In this chapter, our aim is to recover sharp bounds for the true and empirical excess risks of the least-squares estimators of the density, but based on general arguments concerning M-estimation with regular contrast explained in details Chapters 2 and 7. In particular, we avoid the use of explicit formula given in (6.1), (6.2), (6.3), (6.4) and our results could be easily extended to other linear contrasts. Moreover, the bounds that we provide in Section 6.2 are optimal at the first order, and in the case of the empirical excess risk on M, we recover the same magnitude of the deviations bounds as those given by Lerasle in Proposition 2.2.1 of [56]. We do not consider the selection of least-squares estimators of density, as it has already been done in Lerasle [56]. In particular, the author consider all pairs of models of the given collection and this gives a slight improvement of the technology exposed in Arlot and Massart [10].
The chapter is organized as follows. After introducing the precise framework in Section 6.1, we derive in Section 6.2 sharp upper and lower bounds for the true excess risk of the least-squares estimators and its empirical counterpart. We give two theorems, corresponding to different set of assumptions depending of the fact that the unknown density f is of finite sup-norm or simply an element of $L_2(\mu)$, and compare ours results to Lerasle's ones. The proofs are postponed to the end of the Chapter.

6.1 Framework and notations

We assume that we have n i.i.d. observations $(\xi_1, ..., \xi_n)$ with common unknown law P on a measurable space $(\mathcal{Z}, \mathcal{T})$ and that there exists a known probability measure μ on $(\mathcal{Z}, \mathcal{T})$ such that P admits a density f with respect to μ :

$$f = \frac{dP}{d\mu} \ .$$

We endow the space of square integrable measurable functions for the law μ, namely

$$L_2(\mu) = \left\{ s \ , \ \mu\left(s^2\right) < +\infty \right\} \ ,$$

with its natural Hilbertian structure associated to the inner product

$$\langle s, t \rangle = \mu(st) = \int_{\mathcal{Z}} st d\mu$$

and the Hilbertian norm $\|.\|$ is defined by

$$\|s\|^2 = \|s\|^2_{L_2(\mu)} = \langle s, s \rangle = \mu\left(s^2\right) = \int_{\mathcal{Z}} s^2 d\mu \ .$$

We will always assume in the following that f is an element of $L_2(\mu)$ but we will also consider the more restrictive assumption (**H1**) of uniform boundedness of f on $\mathcal{Z}$:

(H1) The unknown density f is uniformly bounded on $\mathcal{Z}$: a positive constant A_∞ exists such that

$$\|f\|_\infty \leq A_\infty < +\infty \ .$$

Moreover, we assume that there exists a given function s_0, typically $s_0 \equiv 1$ if $\mathcal{Z}$ is the unit interval or $s_0 \equiv 0$, and another function s_* such that

$$f = s_0 + s_* \ \text{ and } \int_{\mathcal{Z}} s_* s_0 d\mu = 0 \ .$$

Our goal is to estimate s_*. Considering a generic random variable of law P independent of the sample $(\xi_1, ..., \xi_n)$, we denote expectations in a functional way : for a suitable function f

$$Pf = P(f) = \mathbb{E}[f(\xi)]$$
$$\mu f = \mu(f) = \int_{\mathcal{Z}} f d\mu$$

and if

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$$

denote the empirical distribution associated to the data $(\xi_1, ..., \xi_n)$,

$$P_n f = P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(\xi_i) \ .$$

We also define the orthogonal vector space of s_0 in $L_2(\mu)$, namely

$$\{s_0\}^{\perp} = \{s \in L_2(\mu) \ , \ \langle s, s_0\rangle = 0\} \ .$$

We thus have $s_* \in \{s_0\}^{\perp}$. Now, let $s \in \{s_0\}^{\perp}$, we write

$$\begin{aligned}\|s - s_*\|^2 &= \|s\|^2 - 2\langle s, s_*\rangle + \|s_*\|^2 \\ &= \|s\|^2 - 2\langle s, f\rangle + \|s_*\|^2 \\ &= \|s\|^2 - 2Ps + \|s_*\|^2\end{aligned}$$

and we deduce that

$$\begin{aligned}s_* &= \arg\min_{s\in\{s_0\}^{\perp}} \left\{\|s\|^2 - 2Ps\right\} \\ &= \arg\min_{s\in\{s_0\}^{\perp}} P(Ks) \ ,\end{aligned} \tag{6.5}$$

where the least-squares contrast $K : L_2(\mu) \longrightarrow L_1(P)$ satisfies

$$Ks = \|s\|^2 - 2s \ , \text{ for all } s \in L_2(\mu) \ . \tag{6.6}$$

Now, let us take a finite dimensional vector space $M \subset \{s_0\}^{\perp}$. For every $s \in M$,

$$\langle s, s_0\rangle = \int_{\mathcal{Z}} s s_0 d\mu = 0 \ .$$

The considered estimator on M is the least-squares estimator, defined as follows

$$\begin{aligned}s_n &\in \arg\min_{s\in M} P_n(Ks) \\ &= \arg\min_{s\in M} \left\{\|s\|^2 - 2P_n s\right\} \ .\end{aligned} \tag{6.7}$$

It is easy to check that such an estimator exists, and if $(\varphi_k)_{k=1}^{D}$ is an orthonormal basis of $(M, \|\cdot\|)$,

$$s_n = \sum_{k=1}^{D} P_n(\varphi_k)\varphi_k \ .$$

6.1.1 Excess risk and contrast

As defined in (6.7), s_n is the well-known empirical risk minimizer on M of the least-squares contrast. For any $s \in L_2(\mu)$, the quantity $P(Ks)$ is called the risk of the function s. We notice that for any $s \in \{s_0\}^\perp$,

$$\begin{aligned} P(Ks - Ks_*) &= PKs - PKs_* \\ &= \|s\|^2 - 2\langle s, f\rangle - \|s_*\|^2 + 2\langle s_*, f\rangle \\ &= \|s\|^2 - 2\langle s, s_*\rangle + \|s_*\|^2 \\ &= \|s - s_*\|^2 \geq 0 \ , \end{aligned} \tag{6.8}$$

and so $P(Ks - Ks_*)$, which is called the excess risk of s, is the $L_2(\mu)$ loss. If we denote by s_M the linear projection of s_* onto M in $L_2(\mu)$, we have

$$PKs_M - PKs_* = \inf_{s \in M} \{PKs - PKs_*\} \ , \tag{6.9}$$

and from (6.9), we deduce that

$$s_M = \arg\min_{s \in M} PK(s) \ .$$

We also notice that by the Pythagorean theorem we have for all $s \in M$,

$$\|s - s_*\|^2 = \|s - s_M\|^2 + \|s_M - s_*\|^2$$

and so it holds for all $s \in M$,

$$P(Ks - Ks_M) = \|s - s_M\|^2 \geq 0 \ .$$

Our aim is to study the performance of least-squares estimators, that we measure by their excess risk. We thus look at the random quantity $P(Ks_n - Ks_*)$. Moreover, as we can write

$$P(Ks_n - Ks_*) = P(Ks_n - Ks_M) + P(Ks_M - Ks_*) \ ,$$

we more precisely focus on the quantity

$$P(Ks_n - Ks_M) \geq 0 \ ,$$

that we want to bound with high probability. Abusively, we will often call this last quantity the excess risk of the estimator (on M) or the true excess risk of s_n, by opposition to the empirical excess risk for which the expectation is taken over the empirical measure :

$$P_n(Ks_M - Ks_n) \geq 0 \ .$$

Let us define

$$\begin{aligned} \psi_{1,M}(z) &\equiv -2 \\ \psi_0^s &= \|s\|^2 - \|s_M\|^2 \end{aligned} \tag{6.10}$$

so that we can write for $s \in M$,

$$Ks - Ks_M = \psi_0^s + \psi_{1,M} \cdot (s - s_M) \ . \tag{6.11}$$

In doing so, our aim is to emphasize the fact that in the following analysis, no references are needed to special values of $\psi_{1,M}$ and ψ_0^s. More precisely we only need to assume that

$$\|\psi_{1,M}\|_\infty \leq A_{1,M} < +\infty \tag{6.12}$$

for some positive constant $A_{1,M}$ and (6.12) is automatically satisfied for least-squares density estimation with $A_{1,M} = 2$ using (6.10). Moreover, it makes it easier to relate the situation of this chapter where the contrast is linear to those of chapters 3 and 5 where second orders terms appear in the expansion of the considered contrast.

6.1.2 Linear models

Recall that the model M that we consider is a finite dimensional linear vector space. The linear dimension is written D.
Let us define a function Ψ_M on $\mathcal{Z}$, called the unit envelope, such that

$$\Psi_M(z) = \frac{1}{\sqrt{D}} \sup_{s \in M, \|s\| \leq 1} |s(z)| \ . \tag{6.13}$$

As M is a finite dimensional real vector space, the supremum in (6.13) can also be taken over a countable subset of M, so Ψ_M is a measurable function. The assumption that we make on M is classical, see for example [25] or [13], and rather weak. It states that the unit envelope of M has a finite sup-norm :

(H2) The unit envelope is uniformly bounded from above on $\mathcal{Z}$:

$$\|\Psi_M\|_\infty \leq A_{3,M} < \infty \ .$$

As shown in [13], assumption (**H2**) is satisfied for a very large class of linear models such as some histograms and piecewise polynomials models, models with trigonometric basis or regular wavelet basis, when for example $\mathcal{Z}$ is the unit interval $[0,1]$ and μ is the Lebesgue measure Leb on $\mathcal{Z}$.

6.1.3 Complexity of a linear model M

As we will see in Section 6.2, the rate of convergence of the excess risks on a model M is determined by a quantity that relates the structure of the model to the unknown law P. We call this quantity the complexity of the model M and we denote it by $\mathcal{C}_M$. More precisely, we define

$$\mathcal{C}_M = \frac{1}{4} D \times \mathcal{K}_{1,M}^2$$

where

$$\mathcal{K}_{1,M}^2 = P\left(\psi_{1,M}^2 \cdot \Psi_M^2\right) - \frac{1}{D} \sup_{s \in M, \ \|s\| \leq 1} \left[P\left(\psi_{1,M} \cdot s\right)\right]^2 \ . \tag{6.14}$$

We will see right below that $\mathcal{K}_{1,M}^2$ is indeed nonnegative. The quantity $\mathcal{K}_{1,M} = \sqrt{\mathcal{K}_{1,M}^2} \geq 0$ is called the normalized complexity of the model M.
Let us take an orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|)$. By using Cauchy-Schwarz inequality in (6.13), we have

$$\Psi_M = \sqrt{\frac{1}{D} \sum_{k=1}^D \varphi_k^2} \ . \tag{6.15}$$

Moreover

$$\begin{aligned} \sup_{s \in M, \ \|s\| \leq 1} \left[P\left(\psi_{1,M} \cdot s\right)\right]^2 &= \sup_{(\beta_k)_{k=1}^D \in \mathbb{R}^D, \ \sum \beta_k^2 \leq 1} \left[\sum_{k=1}^D \beta_k P\left(\psi_{1,M} \cdot \varphi_k\right)\right]^2 \\ &= \sum_{k=1}^D \left[P\left(\psi_{1,M} \cdot \varphi_k\right)\right]^2 \ , \end{aligned} \tag{6.16}$$

where the last equality again follows from Cauchy-Schwarz inequality. By combining (6.14), (6.15) and (6.16) we deduce that

$$\mathcal{K}_{1,M}^2 = \frac{1}{D} \sum_{k=1}^D \mathrm{Var}\left(\psi_{1,M} \cdot \varphi_k\right) \geq 0 \ , \tag{6.17}$$

which in particular proves that $\mathcal{K}_{1,M}^2$ is nonnegative.
As we have $\psi_{1,M} \equiv -2$, we deduce from (6.17) that

$$\mathcal{K}_{1,M}^2 = \frac{4}{D} \sum_{k=1}^{D} \mathrm{Var}\left(\varphi_k\right) \tag{6.18}$$

and thus it holds

$$\mathcal{C}_M = \sum_{k=1}^{D} \mathrm{Var}\left(\varphi_k\right) \ .$$

From (6.18) and since for any $k \in \{1, ..., D\}$,

$$\mathrm{Var}\left(\varphi_k\right) \leq P\left(\varphi_k^2\right) = 1 \ ,$$

we can also deduce that

$$\mathcal{K}_{1,M} \leq 2 \ . \tag{6.19}$$

We need the following assumption, ensuring that the normalized complexity $\mathcal{K}_{1,M}$ indeed behaves like a constant.

(H3) Lower bound on the normalized complexity : a positive constant $A_{\mathcal{K},-}$ exists such that

$$\mathcal{K}_{1,M} \geq A_{\mathcal{K},-} > 0 \ . \tag{6.20}$$

Assumption (**H3**) is automatically satisfied for standard finite dimensional linear models such as histograms generated by a finite partition of $\mathcal{Z}$ or more generally, piecewise polynomials generated by a finite partition of $\mathcal{Z}$ and uniformly bounded in their degree. Indeed, consider the model of histograms generated by a finite partition Λ_M of $\mathcal{Z}$. The family $(\varphi_I)_{I \in \Lambda_M}$ defined by

$$\varphi_I : z \in \mathcal{Z} \longmapsto \varphi_I\left(z\right) = \frac{\mathbf{1}_{z \in I}}{\sqrt{\mu\left(I\right)}} \ , \ \text{for all } I \in \Lambda_M \ ,$$

is an orthonormal basis of $(M, \|\cdot\|)$ and it holds

$$\mathcal{K}_{1,M}^2 = \frac{4}{|\Lambda_M|} \sum_{I \in \Lambda_M} \mathrm{Var}\left(\varphi_I\right) = \frac{4}{|\Lambda_M|} \sum_{I \in \Lambda_M} \left(1 - \mu\left(I\right)\right) = 4\left(1 - |\Lambda_M|^{-1}\right) \ .$$

Hence, if the number of elements in the considered partition is larger than two, assumption (**H3**) holds for the model M with $A_{\mathcal{K},-} = \sqrt{2}$. Now, consider more generally some $r \in \mathbb{N}$ and the model M_r of piecewise polynomials generated by Λ_M and of degree less or equal to r. We have $M \subset M_r$ (and $M = M_0$) and $\dim\left(M_r\right) = \left(r+1\right)|\Lambda_M|$, we thus deduce that $\mathcal{K}_{1,M_r}^2 \geq 4\left(r+1\right)^{-1}\left(1 - |\Lambda_M|^{-1}\right)$. Finally, if the number of elements in the considered partition is larger than two, assumption (**H3**) holds for the model M_r with

$$A_{\mathcal{K},-} = \sqrt{\frac{2}{r+1}} > 0 \ .$$

6.2 True and empirical excess risk bounds on a fixed model

In this section, we state upper and lower bounds for the true excess risk on M, $P\left(Ks_n - Ks_M\right)$ and for its empirical counterpart $P_n\left(Ks_M - Ks_n\right)$. We show that under reasonable assumptions the true excess risk is equivalent to the empirical one for a dimension of model not too small. Let us start with the weaker set of assumptions, where we only assume that the unknown density f belongs to $L_2\left(\mu\right)$ and that the model has a unit envelope uniformly bounded, see (**H2**).

Theorem 6.1 *Let $\alpha > 0$ and M a linear model of finite dimension D. Assume that assumptions (**H2**) and (**H3**) hold. If there exists a positive constant A_- such that*

$$n \geq D \geq A_- (\ln n)^3 > 0 \ ,$$

then there exists a positive finite constant A_0, only depending on α, $A_{3,M}$, $\|f\|$ and $A_{\mathcal{K},-}$, such that by setting

$$\varepsilon_n = A_0 \frac{(\ln n)^{1/4}}{D^{1/8}} \ , \tag{6.21}$$

we have for all $n \geq n_0 (A_{3,M}, A_{\mathcal{K},-}, \|f\|, A_-, \alpha)$,

$$\mathbb{P}\left[P(Ks_n - Ks_M) \leq (1-\varepsilon_n)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq 2n^{-\alpha} \ , \tag{6.22}$$

$$\mathbb{P}\left[P(Ks_n - Ks_M) \geq (1+\varepsilon_n)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq 2n^{-\alpha} \ , \tag{6.23}$$

$$\mathbb{P}\left[P_n(Ks_M - Ks_n) \leq \left(1-\varepsilon_n^2\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq n^{-\alpha} \ , \tag{6.24}$$

$$\mathbb{P}\left[P_n(Ks_M - Ks_n) \geq \left(1+\varepsilon_n^2\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq n^{-\alpha} \ . \tag{6.25}$$

In the previous theorem we achieve sharp upper and lower bounds for the true and empirical excess risks on M. They are optimal at the first order since the leading constants are equal for upper and lower bounds. Moreover, although our proofs given in Sections 6.3.2 and 6.3.1 follow from a rather general strategy and can be adapted for the estimation of a regression function or maximum likelihood estimation of density, we recover in inequalities (6.24) and (6.25), concerning the empirical excess risk on M, some results obtained by Lerasle [56] that are based on a different approach relying on explicit formula (6.3), (6.4) and thus only valid for least-squares estimation of density. Recall that in this case, $P(Ks_n - Ks_M) = P_n(Ks_M - Ks_n)$ by (6.4) and so inequalities (6.24) and (6.25) are still valid for the excess risk $P(Ks_n - Ks_M)$. Without using equality (6.4), we derive inequalities (6.22) and (6.23) for the true excess risk on M, that only rely on the linear structure of the least-squares contrast. But in this case, there is a loss in second order terms as the deviations from the first order change from the single for the empirical excess risk to the square root for the true excess risk. We conjecture, due to a comparison with results obtained by Arlot and Massart in [10] in the regression setting, see the discussion in Section 3.4.3 of Chapter 3, that the loss in the deviations occurring for the true excess risk compared the empirical one are quite necessary in general, that is when second order terms appear in the expansion of the considered contrast.

Let us now compare our results to Lerasle's ones. The results of Lerasle [56] related to Theorem 6.1 are exposed in Proposition 2.1 of [56] and can be stated as follows, using some of our notations. Setting $\mathcal{C}_M = \frac{1}{4}D\mathcal{K}_{1,M}^2$ as in Section 6.1.3,

$$B_M = \{s \in M, \|s\| \leq 1\}$$

the unit ball of $(M, \|\cdot\|)$,

$$e_M = \frac{1}{n}\sup_{s\in B_M}\|s\|_\infty^2 \quad \text{and} \quad v_M^2 = \sup_{s\in B_M}\mathrm{Var}(s) \ ,$$

it holds by Proposition 2.1 of [56], for all $x > 0$,

$$\mathbb{P}\left(P(Ks_n - Ks_M) - \frac{\mathcal{C}_M}{n} > \frac{\mathcal{C}_M^{3/4}\left(e_M x^2\right)^{1/4} + 0.7\sqrt{\mathcal{C}_M v_M^2 x} + 0.15 v_M^2 x + e_M x^2}{n}\right) \leq e^{-x/20} \tag{6.26}$$

and

$$\mathbb{P}\left(\frac{\mathcal{C}_M}{n} - P\left(Ks_n - Ks_M\right) > \frac{1.8\mathcal{C}_M^{3/4}\left(e_M x^2\right)^{1/4} + 1.71\sqrt{\mathcal{C}_M v_M^2 x} + 4.06 e_M x^2}{n}\right) \leq 2.8e^{-x/20} \ . \tag{6.27}$$

Take x of order $\ln(n)$ in (6.26) and (6.27). Notice that by (6.19) and (**H3**), $\mathcal{C}_M$ is of the same order as D. Moreover, we have by (**H2**),

$$e_M \leq \frac{DA_{3,M}}{n}$$

and

$$\begin{aligned} v_M^2 &\leq \sup_{s\in B_M} P\left(s^2\right) \\ &= \sup_{s\in B_M} \int s^2 f d\mu \\ &\leq \|f\| \sup_{s\in B_M} \left\|s^2\right\| \qquad \text{by Cauchy-Schwarz inequality} \\ &\leq \|f\| \sup_{s\in B_M} \|s\|_\infty \times \sup_{s\in B_M} \|s\| \\ &\leq \|f\| A_{3,M}\sqrt{D} \qquad \text{by (\textbf{H2}).} \end{aligned}$$

Hence, e_M and v_M^2 are respectively of order Dn^{-1} and $\sqrt{D}$ and we see that the deviations in (6.26) and (6.27) are of order

$$\frac{\sqrt{\ln n}}{D^{1/4}} \cdot \frac{\mathcal{C}_M}{n}$$

as in inequalities (6.24) and (6.25) of Theorem 6.1.
We now turn to some upper and lower bounds for the true and empirical excess risks under the assumption that the unknown density f is uniformly bounded on $\mathcal{Z}$. We thus slightly improve the bounds given in Theorem 6.1, where we only assume that $f \in L_2(\mu)$.

Theorem 6.2 *Let $\alpha > 0$ and M a linear model of finite dimension D. Assume that (**H1**), (**H2**) and (**H3**) hold. If there exists a positive constant A_- such that*

$$n \geq D \geq A_- (\ln n)^2 > 0 \ ,$$

then there exists a positive finite constant A_0, only depending on $A_{3,M}, A_{\mathcal{K},-}, A_\infty$ and α, such that by setting

$$\varepsilon_n = A_0 \max\left\{\left(\frac{\ln n}{D}\right)^{1/4} \ , \ \frac{(\ln n)^{1/4}}{n^{1/8}}\right\} \ ,$$

we have for all $n \geq n_0(A_{3,M}, A_{\mathcal{K},-}, A_\infty, A_-)$,

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \leq (1-\varepsilon_n)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq 2n^{-\alpha}, \tag{6.28}$$

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \geq (1+\varepsilon_n)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq 2n^{-\alpha}, \tag{6.29}$$

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\right) \leq \left(1-\varepsilon_n^2\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq n^{-\alpha}, \tag{6.30}$$

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\right) \geq \left(1+\varepsilon_n^2\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq n^{-\alpha}. \tag{6.31}$$

As we have, for all $1 \leq D \leq n$,

$$\max\left\{\left(\frac{\ln n}{D}\right)^{1/4} \ , \ \frac{(\ln n)^{1/4}}{n^{1/8}}\right\} \leq \frac{(\ln n)^{1/4}}{D^{1/8}} \ ,$$

we deduce that the deviations given in Theorem 6.2 slightly improve those of Theorem 6.1. Again, straightforward computations using inequalities (6.26) and (6.27) given by Lerasle in [56], allow to recover the same magnitude of deviations as in Theorem 6.2 in the case where (**H1**) holds. The remainder of the chapter is dedicated to the proofs of Theorems 6.1 and 6.2.

6.3 Proofs

6.3.1 Proofs of the theorems

In order to express the quantities of interest, we need preliminaries definitions.
Elements of an orthonormal basis in $(M, \|\cdot\|)$ are denoted by φ_k, $k = 1, ..., D$. Let us define several slices of excess risk on the model M :

$$\mathcal{F}_C = \{s \in M, P(Ks - Ks_M) \leq C\}$$
$$\mathcal{F}_{>C} = \{s \in M, P(Ks - Ks_M) > C\}$$

and for any interval $I \subset \mathbb{R}_+$,

$$\mathcal{F}_I = \{s \in M, P(Ks - Ks_M) \in I\} \ .$$

We also define for any $L > 0$,

$$D_L = \{s \in M, P(Ks - Ks_M) = L\} \ .$$

The proof of Theorem 6.1 relies on Lemmas 6.5 and 6.7 stated in Section 6.3.2 below, and that give sharp estimates of suprema of the empirical process indexed by the constrasted functions over several slices of interest.

Proof of Theorem 6.1. Let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $(M, \|\cdot\|)$. We divide the proof of Theorem **6.1** in four parts, corresponding to the four Inequalities (6.22), (6.23), (6.24) and (6.25). The values of A_0 defined in (6.21) will be then fixed at the end of the proof.

Proof of Inequality (6.22). Let $r \in (1, 2]$ to be fixed later and $C > 0$ such that

$$rC = \frac{D}{4n}\mathcal{K}_{1,M}^2 \ . \tag{6.32}$$

It holds

$$\begin{aligned}
&\mathbb{P}(P(Ks_n - Ks_M) \leq C) \\
&\leq \mathbb{P}\left(\inf_{s\in\mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s\in\mathcal{F}_{>C}} P_n(Ks - Ks_M)\right) \\
&\leq \mathbb{P}\left(\inf_{s\in\mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s\in\mathcal{F}_{(C,rC]}} P_n(Ks - Ks_M)\right) \\
&= \mathbb{P}\left(\sup_{s\in\mathcal{F}_C} P_n(Ks_M - Ks) \geq \sup_{s\in\mathcal{F}_{(C,rC]}} P_n(Ks_M - Ks)\right) \ .
\end{aligned} \tag{6.33}$$

Now, by (6.32) we have $C \leq \frac{1}{4}(1+\tau_n)^2 \frac{D}{n}\mathcal{K}_{1,M}^2$ where $\tau_n = L_{A_{\mathcal{K},-},A_{3,M},\|f\|,\beta} \times \frac{\sqrt{\ln n}}{D^{1/4}}$ is defined in Lemma 6.1, so we can apply Lemma 6.5 with $\alpha = \beta$ and it holds,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C} P_n\left(Ks_M - Ks\right) \geq (1+\tau_n)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq n^{-\alpha}\ . \tag{6.34}$$

Moreover, by using (**H3**) and (6.19) in (6.32) we get

$$\frac{D}{4n}A_{\mathcal{K},-}^2 \leq rC \leq \frac{D}{n}\ .$$

We then apply Lemma 6.7 with

$$\alpha = \beta,\ A_l = A_{\mathcal{K},-}^2/4,\ A_u = 1$$

so it holds for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, \|f\|, A_-\right)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},\|f\|,\alpha} \times \frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq n^{-\alpha}\ . \tag{6.35}$$

Now, from (6.34) and (6.35) we deduce that a positive constant $\tilde{A}_0$ exists, only depending on $A_{3,M}, A_{\mathcal{K},-}, \|f\|$ and α, such that for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, \|f\|, A_-\right)$, it holds on the same event of probability at least $1 - 2n^{-\alpha}$,

$$\sup_{s\in\mathcal{F}_C} P_n\left(Ks_M - Ks\right) \leq \left(1 + \tilde{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C \tag{6.36}$$

and

$$\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \geq \left(1 - \tilde{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\ . \tag{6.37}$$

Hence, from (6.36) and (6.7) we deduce, using (6.33), that if we choose $r \in (1,2]$ such that

$$\left(1 + \tilde{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C < \left(1 - \tilde{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC \tag{6.38}$$

then, for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, \|f\|, A_-\right)$, $P\left(Ks_n - Ks_M\right) \geq C$ with probability at least $1 - 2n^{-\alpha}$. Now, by (6.32) it holds

$$\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} = 2rC = \frac{1}{2}\frac{D}{n}\mathcal{K}_{1,M}^2\ ,$$

and by consequence inequality (6.38) is equivalent to

$$\left(1 - 2\tilde{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)r - 2\left(1 + \tilde{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{r} + 1 > 0\ . \tag{6.39}$$

Moreover, since $D \geq A_-\left(\ln n\right)^3$ we have for all $n \geq n_0\left(A_-, \tilde{A}_0\right)$,

$$\tilde{A}_0\frac{\sqrt{\ln n}}{D^{1/4}} \leq \frac{1}{4} \tag{6.40}$$

and so, for all $n \geq n_0\left(A_-, \tilde{A}_0\right)$, simple computations using (6.40) show that by taking

$$r = 1 + 48\sqrt{\tilde{A}_0}\frac{(\ln n)^{1/4}}{D^{1/8}} \tag{6.41}$$

inequality (6.39) is satisfied. Notice that, for all $n \geq n_0\left(A_-, \tilde{A}_0\right)$, $0 < 48\sqrt{\tilde{A}_0}\frac{(\ln n)^{1/4}}{D^{1/8}} < 1$, so that $r \in (1,2)$. Finally, we compute C by (6.32) and (6.41). For all $n \geq n_0\left(A_-, \tilde{A}_0\right)$,

$$C = \frac{rC}{r} = \frac{1}{1 + 48\sqrt{\tilde{A}_0}\left(\frac{(\ln n)^{1/4}}{D^{1/8}}\right)}\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \geq \left(1 - 48\sqrt{\tilde{A}_0}\frac{(\ln n)^{1/4}}{D^{1/8}}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 > 0\ , \quad (6.42)$$

which yields the result.

Proof of Inequality (6.23). Let $C > 0$ and $\delta \in \left(0, \frac{1}{2}\right)$ to be fixed later satisfying

$$(1-\delta)\, C = \frac{D}{4n}\mathcal{K}_{1,M}^2 \quad (6.43)$$

and

$$C \geq \frac{1}{4}\left(1 + \tau_n\right)^2 \frac{D}{n}\mathcal{K}_{1,M}^2\ , \quad (6.44)$$

where $\tau_n = L_{A_{\mathcal{K},-},A_{3,M},\|f\|,\beta} \times \frac{\sqrt{\ln n}}{D^{1/4}}$ is defined in Lemma 6.1. We have

$$\begin{aligned}
&\mathbb{P}\left(P\left(Ks_n - Ks_M\right) > C\right) \\
&\leq \mathbb{P}\left(\inf_{s\in\mathcal{F}_C} P_n\left(Ks - Ks_M\right) \geq \inf_{s\in\mathcal{F}_{>C}} P_n\left(Ks - Ks_M\right)\right) \\
&= \mathbb{P}\left(\sup_{s\in\mathcal{F}_C} P_n\left(Ks_M - Ks\right) \leq \sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right)\right) \\
&\leq \mathbb{P}\left(\sup_{s\in\mathcal{F}_{\left[\frac{C}{2},(1-\delta)C\right]}} P_n\left(Ks_M - Ks\right) \leq \sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right)\right)\ . \quad (6.45)
\end{aligned}$$

Now by (6.44) we apply Lemma 6.5 with $\alpha = \beta$ and we get

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \left(1+\tau_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq n^{-\alpha}\ . \quad (6.46)$$

Moreover, by (6.43), (**H3**) and (6.19) we can apply Lemma 6.7 with the constant C in Lemma 6.7 replaced by $C/2$, $\alpha = \beta$, $r = 2\left(1-\delta\right)$, $A_u = 1$, $A_l = A_{\mathcal{K},-}^2/4$ and so it holds, for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, \|f\|, A_-\right)$,

$$\mathbb{P}\left(\begin{array}{c}\sup_{s\in\mathcal{F}_{\left[\frac{C}{2},(1-\delta)C\right]}} P_n\left(Ks_M - Ks\right) \\ \leq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|,\alpha} \times \frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{(1-\delta)CD}{n}}\mathcal{K}_{1,M} - (1-\delta)\, C\end{array}\right) \leq n^{-\alpha}\ . \quad (6.47)$$

Hence, from (6.46) and (6.47), we deduce that a positive constant $\check{A}_0$ exists, only depending on $A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|$ and α, such that for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, \|f\|, A_-\right)$ it holds on the same event of probability at least $1 - 2n^{-\alpha}$,

$$\sup_{s\in\mathcal{F}_{\left[\frac{C}{2},(1-\delta)C\right]}} P_n\left(Ks_M - Ks\right) \geq \left(1 - \check{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{(1-\delta)\, CD}{n}}\mathcal{K}_{1,M} - (1-\delta)\, C \quad (6.48)$$

and

$$\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \leq \left(1 + \check{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\ . \tag{6.49}$$

Now, from (6.48), (6.49) and (6.45), we deduce that if we choose $\delta \in \left(0, \frac{1}{2}\right)$ such that (6.44) and

$$\left(1 + \check{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C < \left(1 - \check{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{(1-\delta)\,CD}{n}}\mathcal{K}_{1,M} - (1-\delta)\,C \tag{6.50}$$

are satisfied then, for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, \|f\|, A_-\right)$, $P\left(Ks_n - Ks_M\right) \leq C$ with probability at least $1 - 2n^{-\alpha}$. By (6.43) it holds

$$\sqrt{\frac{(1-\delta)\,CD}{n}}\mathcal{K}_{1,M} = 2\,(1-\delta)\,C = \frac{1}{2}\frac{D}{n}\mathcal{K}_{1,M}^2\ ,$$

and by consequence inequality (6.50) is equivalent to

$$\left(1 - 2\check{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)(1-\delta) - 2\left(1 + \check{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{1-\delta} + 1 > 0\ . \tag{6.51}$$

Moreover, since $D \geq A_-\left(\ln n\right)^3$ we have for all $n \geq n_0\left(A_{\mathcal{K},-}, A_{3,M}, \|f\|, A_-, \alpha\right)$,

$$\max\left\{\check{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\ ;\ \tau_n\right\} < \frac{1}{72} \tag{6.52}$$

and so, for all $n \geq n_0\left(A_{\mathcal{K},-}, A_{3,M}, \|f\|, A_-, \alpha\right)$, simple computations show that by taking

$$\delta = 6\max\left\{\check{A}_0\frac{(\ln n)^{1/4}}{D^{1/8}}\ ;\ \sqrt{\tau_n}\right\} =: 6\check{A}_0\frac{(\ln n)^{1/4}}{D^{1/8}} \tag{6.53}$$

for a positive constant $\check{A}_0$ depending only on $A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|$ and α. Hence, inequalities (6.51) and (6.44) are satisfied and $\delta \in \left(0, \frac{1}{2}\right)$. Finally, we compute C by (6.43) and (6.53), for all $n \geq n_0\left(A_{\mathcal{K},-}, A_{3,M}, \|f\|, A_-, \alpha\right)$,

$$0 < C = \frac{(1-\delta)\,C}{(1-\delta)} = \frac{1}{(1-\delta)}\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 \leq \left(1 + 12\check{A}_0\frac{(\ln n)^{1/4}}{D^{1/8}}\right)\frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\ , \tag{6.54}$$

which readily yields the result.

Proof of Inequality (6.24). Let $C = \frac{D}{8n}\mathcal{K}_{1,M}^2 > 0$ and let $r = 2$. By (**H3**) and (6.19) we have

$$\frac{D}{4n}A_{\mathcal{K},-}^2 \leq rC = \frac{D}{4n}\mathcal{K}_{1,M}^2 \leq \frac{D}{n}$$

so we apply Lemma 6.7 with $\alpha = \beta$, $A_l = A_{\mathcal{K},-}^2/4$ and $A_u = 1$. Hence, it holds for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|, A_-\right)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|,\alpha} \times \frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq n^{-\alpha}\ , \tag{6.55}$$

and as $rC = \frac{D}{4n}\mathcal{K}_{1,M}^2$, if we set $\hat{A}_0 = 2L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|,\alpha}$ with $L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|,\alpha}$ the constant appearing in (6.55), we get

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - \hat{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\frac{D}{4n}\mathcal{K}_{1,M}^2\right) \leq n^{-\alpha} \ . \tag{6.56}$$

Notice that

$$P_n\left(Ks_M - Ks_n\right) = \sup_{s\in M} P_n\left(Ks_M - Ks\right) \geq \sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \ ,$$

so from (6.56) we deduce that

$$\mathbb{P}\left(P_n\left(Ks_M - Ks_n\right) \geq \left(1 - \hat{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\frac{D}{4n}\mathcal{K}_{1,M}^2\right) \geq 1 - n^{-\alpha} \ . \tag{6.57}$$

Proof of Inequality (6.25). Let

$$C = \frac{1}{4}\left(1+\tau_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2 > 0 \ , \tag{6.58}$$

where τ_n is defined in Lemma 3.13 applied with $\beta = \alpha$. By (6.82) of Lemma 6.5 applied with $\alpha = \beta$ it holds,

$$\mathbb{P}\left(\sup_{s\in M} P_n\left(Ks_M - Ks\right) > C\right) \leq n^{-\alpha} \ ,$$

which gives the result since $\sup_{s\in M} P_n\left(Ks_M - Ks\right) = P_n\left(Ks_M - Ks_n\right)$ and since a positive constant $\bar{A}_0$ exists, only depending on $A_{\mathcal{K},-}, A_{3,M}, \|f\|$ and α, such that

$$C \leq \frac{1}{4}\left(1 + \bar{A}_0\frac{\sqrt{\ln n}}{D^{1/4}}\right)\frac{D}{n}\mathcal{K}_{1,M}^2 \ . \tag{6.59}$$

Conclusion. To complete the proof of Theorem 3.1, just notice that by (6.42), (6.54), (6.57) and (6.59),

$$A_0 = \max\left\{48\sqrt{\tilde{A}_0},\ 12\check{A}_0,\ \sqrt{\hat{A}_0},\ \sqrt{\bar{A}_0}\right\}$$

is convenient. ■

The proof of Theorem 6.2 follows from a straightforward adaptation of the proof of Theorem 6.1. Just replace the use of Lemmas 6.5 and 6.7 by Lemmas 6.6 and 6.35 stated in Section 6.3.2.

6.3.2 Technical lemmas

We provide here with some lemmas needed in the proofs stated in the previous section.

Lemma 6.1 *Let $\beta > 0$. Assume that (**H2**) and (**H3**) hold. Then a positive constant $L_{A_{\mathcal{K},-},A_{3,M},\|f\|,\beta}$ exists, such that by setting*

$$\tau_n = L_{A_{\mathcal{K},-},A_{3,M},\|f\|,\beta} \frac{\sqrt{\ln n}}{D^{1/4}} \ ,$$

we have, for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|)$,

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D} \left(P_n - P\right)^2 \left(\psi_{1,M} \cdot \varphi_k\right)} \geq (1+\tau_n)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} \ .$$

Proof. By Cauchy-Schwarz inequality we have,

$$\chi := \sqrt{\sum_{k=1}^{D} \left(P_n - P\right)^2 \left(\psi_{1,M} \cdot \varphi_k\right)} = \sup\left\{\left|\left(P_n - P\right)\left(\psi_{1,M} \cdot s\right)\right| \ ; \ s \in L_2(\mu) \ \ \& \ \ \|s\| \leq 1\right\} \ .$$

Hence, we get by Bousquet's inequality (7.48), for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi \geq \sqrt{2\sigma^2 \frac{x}{n}} + (1+\delta)\,\mathbb{E}\left[\chi\right] + \left(\frac{1}{3} + \frac{1}{\delta}\right)\frac{bx}{n}\right] \leq \exp\left(-x\right) \ , \tag{6.60}$$

where

$$\begin{aligned}
\sigma^2 &\leq \sup_{\|s\|\leq 1} P\left(\left(\psi_{1,M} \cdot s\right)^2\right) \\
&\leq 4 \sup_{\|s\|\leq 1} \int s^2 f d\mu \\
&\leq 4\|f\| \sup_{\|s\|\leq 1} \left\|s^2\right\| \qquad \text{by Cauchy-Schwarz inequality} \\
&\leq 4\|f\| \sup_{\|s\|\leq 1} \|s\|_\infty \times \sup_{\|s\|\leq 1} \|s\| \\
&\leq 4A_{3,M}\|f\|\sqrt{D} \qquad \text{by } (\mathbf{H2})
\end{aligned} \tag{6.61}$$

and

$$b \leq \sup_{\|s\|_2\leq 1} \|\psi_{1,M} \cdot s - P\left(\psi_{1,M} \cdot s\right)\|_\infty \leq 2 \sup_{\|s\|_2\leq 1} \|\psi_{1,M} \cdot s\|_\infty \leq 4\sqrt{D}A_{3,M} \ \text{ by } (\mathbf{H2}). \tag{6.62}$$

Moreover,

$$\mathbb{E}\left[\chi\right] \leq \sqrt{\mathbb{E}\left[\chi^2\right]} = \sqrt{\frac{D}{n}}\mathcal{K}_{1,M} \ . \tag{6.63}$$

So, by combining (6.61), (6.62) and (6.63) with (6.60), it follows that for all $x > 0$,

$$\mathbb{P}\left[\chi \geq \sqrt{8\|f\|A_{3,M}\sqrt{D}\frac{x}{n}} + (1+\delta)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M} + \left(\frac{1}{3} + \frac{1}{\delta}\right)\frac{4\sqrt{D}A_{3,M}x}{n}\right] \leq \exp\left(-x\right) \ .$$

Hence, taking $x = \beta \ln n$ and $\delta = \frac{\sqrt{\ln n}}{n^{1/4}} \leq \frac{\sqrt{\ln n}}{D^{1/4}}$ gives the result. ■

Lemma 6.2 *Let $\beta > 0$. Assume that (**H1**), (**H2**) and (**H3**) hold. Then a positive constant $L_{A_{\mathcal{K},-},A_{3,M},A_\infty,\beta}$ exists, such that by setting*

$$\tau_n^\infty = L_{A_{\mathcal{K},-},A_{3,M},A_\infty,\beta}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right) ,$$

we have, for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $(M, \|\cdot\|)$,

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \geq (1+\tau_n^\infty)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} .$$

Proof. By Cauchy-Schwarz inequality we have

$$\chi := \sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} = \sup\left\{\left|(P_n - P)(\psi_{1,M} \cdot s)\right|,\ s \in M \ \&\ \|s\| \leq 1\right\} .$$

Hence, we get by Bousquet's inequality (7.48), for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi \geq \sqrt{2\sigma^2\frac{x}{n}} + (1+\delta)\,\mathbb{E}[\chi] + \left(\frac{1}{3}+\frac{1}{\delta}\right)\frac{bx}{n}\right] \leq \exp(-x) , \tag{6.64}$$

where

$$\begin{aligned}\sigma^2 &\leq \sup_{\|s\|\leq 1} P\left((\psi_{1,M}\cdot s)^2\right)\\ &\leq 4\sup_{\|s\|\leq 1}\int s^2 f d\mu\\ &\leq 4A_\infty \sup_{\|s\|\leq 1}\|s\| \leq 4A_\infty\end{aligned} \tag{6.65}$$

and

$$b \leq \sup_{\|s\|_2\leq 1}\|\psi_{1,M}\cdot s - P(\psi_{1,M}\cdot s)\|_\infty \leq 4\sqrt{D}A_{3,M} . \tag{6.66}$$

Moreover,

$$\mathbb{E}[\chi] \leq \sqrt{\mathbb{E}[\chi^2]} = \sqrt{\frac{D}{n}}\mathcal{K}_{1,M} . \tag{6.67}$$

So, by combining (6.65), (6.66) and (6.67) with (6.60), it follows that

$$\mathbb{P}\left[\chi \geq \sqrt{8A_\infty\frac{x}{n}} + (1+\delta)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M} + \left(\frac{1}{3}+\frac{1}{\delta}\right)\frac{4\sqrt{D}A_{3,M}x}{n}\right] \leq \exp(-x) .$$

Hence, taking $x = \beta\ln n$ and $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ gives the result. ■

In the next lemma, we state sharp lower bounds for the mean of the supremum of the empirical process on functions of M belonging to a slice of excess risk. The unknown density f is only assumed to belong to $L_2(\mu)$.

Lemma 6.3 *Let $r > 1$ and $C > 0$. Assume (**H1**) and (**H3**). If positive constants A_-, A_l and A_u exist such that*

$$n \geq D \geq A_-\ln n \quad \text{and} \quad A_l\frac{D}{n} \leq rC \leq A_u\frac{D}{n} ,$$

then for all $n \geq n_0(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|, A_-)$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}}(P_n - P)(\psi_{1,M}\cdot(s_M - s))\right] \geq \left(1 - \frac{L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} > 0 .$$

(6.68)

Proof of Lemma 6.3. First observe that $s \in \mathcal{F}_{(C,rC]}$ implies that $2s_M - s \in \mathcal{F}_{(C,rC]}$, so that

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right] = \mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} \left|(P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right|\right] .$$

In the next step, we apply Corollary 7.2. More precisely, using notations of Corollary 7.2, we set

$$\begin{aligned}
\mathcal{F} &= \left\{\psi_{1,M}\cdot(s_M - s)\,, s \in \mathcal{F}_{(C,rC]}\right\} \ , \\
\varkappa_n^2 &= 4\frac{A_{3,M}}{A_{\mathcal{K},-}} \max\left\{\sqrt{\frac{A_u}{A_l}}\frac{1}{\sqrt{n}}\ ;\ \frac{A_u\,\|f\|}{A_l}\frac{1}{\sqrt{D}}\right\} &(6.69)\\
&\le L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|} D^{-1/2} &(6.70)
\end{aligned}$$

and

$$Z = \sup_{s\in\mathcal{F}_{(C,rC]}} \left|(P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right| \ .$$

We readily get by Cauchy-Schwarz inequality, using (**H3**),

$$\sqrt{\mathbb{E}\left[Z^2\right]} = \sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \ge \sqrt{A_l}\mathcal{K}_{1,M}\frac{D}{n} \ge \sqrt{A_l}A_{\mathcal{K},-}\frac{D}{n} \ . \tag{6.71}$$

Now, as we have

$$\sup_{f\in\mathcal{F}} \|f - Pf\|_\infty \le 2 \sup_{s\in\mathcal{F}_{(C,rC]}} \|\psi_{1,M}\cdot(s_M - s)\|_\infty \le 4\sqrt{rCD}A_{3,M} \quad \text{by (\textbf{H2})},$$

we set $b = 4\sqrt{rCD}A_{3,M}$, and it holds from (6.69) and (6.71),

$$\varkappa_n^2\sqrt{\mathbb{E}\left[Z^2\right]} \ge \frac{b}{n} \ . \tag{6.72}$$

Moreover, we have

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \mathrm{Var}\,(f) &\le \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(\psi_{1,M}\cdot(s_M - s)\right)^2 \\
&\le 4 \sup_{s\in\mathcal{F}_{(C,rC]}} \int (s_M - s)^2 f d\mu \\
&\le 4\,\|f\| \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|(s_M - s)^2\right\| \qquad \text{by Cauchy-Schwarz inequality} \\
&\le 4\,\|f\| \sup_{s\in\mathcal{F}_{(C,rC]}} \|s - s_M\|_\infty \times \sup_{s\in\mathcal{F}_{(C,rC]}} \|s - s_M\| \\
&\le 4A_{3,M}\,\|f\|\, rC\sqrt{D} \qquad \text{by (\textbf{H2})} \ .
\end{aligned}$$

Hence, we take $\sigma^2 = 4A_{3,M}\,\|f\|\,rC\sqrt{D}$ and we get by (6.69) and (6.71),

$$\varkappa_n^2\mathbb{E}\left[Z^2\right] \ge \frac{\sigma^2}{n} \ . \tag{6.73}$$

Finally, since $D \ge A_- \ln n$, we have for all $n \ge n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|\right)$,

$$0 \le \varkappa_n \le \frac{L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|}}{D^{1/4}} < 1$$

and so, using (6.72) and (6.73), we deduce from Corollary 7.2 that for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|\right)$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right] \\
&\geq (1-\varkappa_n A_{1,-})\,\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right)^2 \\
&\geq \left(1 - \frac{L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \ > 0,
\end{aligned}$$

which yields the result. ■

We give in the following lemma the same type of result as in Lemma 6.3, but in the more restrictive assumption where the unknown density is of finite sup-norm.

Lemma 6.4 *Let $r > 1$ and $C > 0$. Assume **(H1)**, **(H2)** and **(H3)**. If positive constants A_-, A_l and A_u exist such that*

$$n \geq D \geq A_- \ln n \quad \textit{and} \quad A_l\frac{D}{n} \leq rC \leq A_u\frac{D}{n}\ ,$$

then for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, A_\infty, A_-\right)$,

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right] \geq \left(1 - \frac{L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty}}{n^{1/4}\wedge D^{1/2}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} > 0\ . \tag{6.74}$$

Proof of Lemma 6.4. First observe that $s \in \mathcal{F}_{(C,rC]}$ implies that $2s_M - s \in \mathcal{F}_{(C,rC]}$, so that

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right] = \mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} \left|(P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right|\right]\ .$$

In the next step, we apply Corollary 7.2. More precisely, using notations of Corollary 7.2, we set

$$\begin{aligned}
\mathcal{F} &= \left\{\psi_{1,M}\cdot(s_M - s)\,, s\in\mathcal{F}_{(C,rC]}\right\}\ , \\
\varkappa_n^2 &= \frac{4}{A_{\mathcal{K},-}}\max\left\{A_{3,M}\sqrt{\frac{A_u}{A_l}}\frac{1}{\sqrt{n}}\ ;\ \frac{A_\infty A_u}{A_l}\frac{1}{D}\right\} &(6.75)\\
&\leq L_{r,A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty}\max\left\{n^{-1/2}\ ;\ D^{-1}\right\} &(6.76)
\end{aligned}$$

and

$$Z = \sup_{s\in\mathcal{F}_{(C,rC]}} \left|(P_n - P)\left(\psi_{1,M}\cdot(s_M - s)\right)\right|\ .$$

We get by Cauchy-Schwarz inequality, using (**H3**),

$$\sqrt{\mathbb{E}\left[Z^2\right]} = \sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \geq \sqrt{A_l}\mathcal{K}_{1,M}\frac{D}{n} \geq \sqrt{A_l}A_{\mathcal{K},-}\frac{D}{n}\ . \tag{6.77}$$

Now, as we have

$$\sup_{f\in\mathcal{F}}\|f - Pf\|_\infty \leq 2\sup_{s\in\mathcal{F}_{(C,rC]}}\|\psi_{1,M}\cdot(s_M - s)\|_\infty \leq 4\sqrt{rCD}A_{3,M} \ \text{ by } (\mathbf{H2})$$

we set $b = 4\sqrt{rCD}A_{3,M}$, and it holds from (6.75) and (6.77),

$$\varkappa_n^2 \sqrt{\mathbb{E}\left[Z^2\right]} \geq \frac{b}{n} \ . \tag{6.78}$$

Moreover, we have

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \mathrm{Var}\left(f\right) &\leq \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)^2 \\
&\leq 4 \sup_{s\in\mathcal{F}_{(C,rC]}} \int \left(s_M - s\right)^2 f d\mu \\
&\leq 4A_\infty \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|s_M - s\right\|^2 \qquad\qquad \text{by } (\mathbf{H1}) \\
&\leq 4A_\infty rC
\end{aligned}$$

Hence, we take $\sigma^2 = 4A_\infty rC$ and we get by (6.75) and (6.77),

$$\varkappa_n^2 \mathbb{E}\left[Z^2\right] \geq \frac{\sigma^2}{n} \ . \tag{6.79}$$

Finally, since $D \geq A_- \ln n$, we have for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, A_\infty\right)$,

$$0 \leq \varkappa_n \leq L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty} \max\left\{n^{-1/4}\ ;\ D^{-1/2}\right\} < 1$$

and so, using (6.78) and (6.79), we deduce from Corollary 7.2 that for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, A_\infty\right)$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right] \\
&\geq \left(1 - \varkappa_n A_{1,-}\right)\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right)^2 \\
&\geq \left(1 - \frac{L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty}}{n^{1/4}\wedge D^{1/2}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \ ,
\end{aligned}$$

which yields the result. ■

In the following lemma we give sharp upper bounds for the supremum of the empirical excess risk on the slides of interest in the case where (**H2**) and (**H3**) hold.

Lemma 6.5 *Let $\beta > 0$ and $C \geq 0$. Assume that (**H2**) and (**H3**). If $C \leq \frac{1}{4}\left(1+\tau_n\right)^2 \frac{D}{n}\mathcal{K}_{1,M}^2$ where $\tau_n = L_{A_{\mathcal{K},-},A_{3,M},\|f\|,\beta} \times \frac{\sqrt{\ln n}}{D^{1/4}}$ is defined in Lemma 6.1, then it holds*

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C} P_n\left(Ks_M - Ks\right) \geq \left(1+\tau_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq n^{-\beta} \ . \tag{6.80}$$

If $C \geq \frac{1}{4}\left(1+\tau_n\right)^2 \frac{D}{n}\mathcal{K}_{1,M}^2$ then it holds

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \left(1+\tau_n\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq n^{-\beta} \ . \tag{6.81}$$

Moreover, we have

$$\mathbb{P}\left[\sup_{s\in M} P_n\left(Ks_M - Ks\right) \geq \frac{1}{4}\left(1+\tau_n\right)^2 \frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq n^{-\beta} \ . \tag{6.82}$$

Proof. Start with

$$\begin{aligned}\sup_{s\in\mathcal{F}_C} P_n\left(Ks_M - Ks\right) &= \sup_{s\in\mathcal{F}_C}\left\{P_n\left(\psi_{1,M}\cdot(s_M-s)\right)-\psi_0^s\right\}\\ &= \sup_{s\in\mathcal{F}_C}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot(s_M-s)\right)-P\left(Ks-Ks_M\right)\right\}\ .\end{aligned}$$

Next, recall that by definition

$$D_L = \left\{s\in M, P\left(Ks-Ks_M\right)=L\right\}\ ,$$

so we have

$$\begin{aligned}&\sup_{s\in\mathcal{F}_C}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot(s_M-s)\right)-P\left(Ks-Ks_M\right)\right\}\\ &= \sup_{0\le L\le C}\ \sup_{D_L}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot(s_M-s)\right)-L\right\}\\ &= \sup_{0\le L\le C}\left\{\sqrt{L}\sqrt{\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)}-L\right\}\ ,\end{aligned}$$

where the last bound follows from Cauchy-Schwarz inequality. Hence, we deduce from Lemma 6.1 that

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot(s_M-s)\right)-P\left(Ks-Ks_M\right)\right\}\ge\sup_{L\le C}\left\{\sqrt{L}\left(1+\tau_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-L\right\}\right]\le n^{-\beta}\ .$$

So, as $C\le\frac{1}{4}\left(1+\tau_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2$ we get by simple calculations that

$$\sup_{0\le L\le C}\left\{\sqrt{L}\left(1+\tau_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-L\right\}=\sqrt{C}\left(1+\tau_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-C$$

and by consequence,

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}\left\{\left(P_n-P\right)\left(\psi_{1,M}\cdot(s_M-s)\right)-P\left(Ks-Ks_M\right)\right\}\ge\sqrt{C}\left(1+\tau_n\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}-C\right]\le n^{-\beta}\ ,$$

which yields (6.80). Now inequality (6.81) follows from the same type of arguments. Inequality (6.82) is a straightforward corollary of (6.80) and (6.81) applied with $C=\frac{1}{4}\left(1+\tau_n\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2$, since we have

$$\sup_{s\in M}P_n\left(Ks_M-Ks\right)=\sup_{s\in\mathcal{F}_C}P_n\left(Ks_M-Ks\right)\vee\sup_{s\in\mathcal{F}_{>C}}P_n\left(Ks_M-Ks\right)\ .$$

■

In the case where (**H1**) hold, we have the following result.

Lemma 6.6 *Let $\beta>0$ and $C\ge 0$. Assume that (**H1**), (**H2**) and (**H3**) hold. If $C\le\frac{1}{4}\left(1+\tau_n^\infty\right)^2\frac{D}{n}\mathcal{K}_{1,M}^2$ where $\tau_n^\infty=L_{A_{\mathcal{K},-},A_{3,M},A_\infty,\beta}\left(\sqrt{\frac{\ln n}{D}}\vee\frac{\sqrt{\ln n}}{n^{1/4}}\right)$ is defined in Lemma 6.2, then it holds*

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_C}P_n\left(Ks_M-Ks\right)\ge\left(1+\tau_n^\infty\right)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M}-C\right]\le n^{-\beta}\ .\tag{6.83}$$

If $C \geq \frac{1}{4}(1+\tau_n)^2 \frac{D}{n}\mathcal{K}_{1,M}^2$ *then it holds*

$$\mathbb{P}\left[\sup_{s\in\mathcal{F}_{>C}} P_n(Ks_M - Ks) \geq (1+\tau_n^\infty)\sqrt{\frac{CD}{n}}\mathcal{K}_{1,M} - C\right] \leq n^{-\beta} . \tag{6.84}$$

Moreover, we have

$$\mathbb{P}\left[\sup_{s\in M} P_n(Ks_M - Ks) \geq \frac{1}{4}(1+\tau_n^\infty)^2 \frac{D}{n}\mathcal{K}_{1,M}^2\right] \leq n^{-\beta} . \tag{6.85}$$

The proof of Lemma 6.6 is similar to the proof of Lemma 6.5. Just replace the use of Lemma 6.1 and the related quantity τ_n given in the proof of Lemma 6.5 by the use of corresponding Lemma 6.2 and related quantity τ_n^∞ in the case of Lemma 6.6 where (**H1**) hold.
In the following lemma, we give a sharp bound for the supremum of the empirical excess risk on a slide of interest in the case where (**H2**) and (**H3**) hold.

Lemma 6.7 *Let* $r > 1$ *and* $C, \beta > 0$. *Assume that (**H2**) and (**H3**) hold. If positive constants* A_-, A_l, A_u *exist such that*

$$n \geq D \geq A_- (\ln n)^3 \quad \text{and} \quad A_l \frac{D}{n} \leq rC \leq A_u \frac{D}{n} ,$$

then a positive constant $L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|,\beta}$ *exists such that, for all* $n \geq n_0(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|, A_-)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n(Ks_M - Ks) \leq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|,\beta} \times \frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq n^{-\beta} .$$

Proof. Start with

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_{(C,rC]}} P_n(Ks_M - Ks) \\
&= \sup_{s\in\mathcal{F}_{(C,rC]}} \{(P_n - P)(Ks_M - Ks) + P(Ks_M - Ks)\} \\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M}\cdot(s_M - s)) - \sup_{s\in\mathcal{F}_{(C,rC]}} P(Ks - Ks_M) \\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M}\cdot(s_M - s)) - rC
\end{aligned} \tag{6.86}$$

and set

$$\begin{aligned}
S_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M}\cdot(s_M - s)) \\
M_{1,r,C} &= \mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} (P_n - P)(\psi_{1,M}\cdot(s_M - s))\right] \\
b_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \|\psi_{1,M}\cdot(s_M - s) - P\psi_{1,M}\cdot(s_M - s)\|_\infty \\
\sigma_{1,r,C}^2 &= \sup_{s\in\mathcal{F}_{(C,rC]}} \mathrm{Var}(\psi_{1,M}\cdot(s_M - s)) .
\end{aligned}$$

By Klein-Rio's Inequality (7.50), we get, for all $\delta, x > 0$,

$$\mathbb{P}\left(S_{1,r,C} \leq (1-\delta)M_{1,r,C} - \sqrt{\frac{2\sigma_{1,r,C}^2 x}{n}} - \left(1 + \frac{1}{\delta}\right)\frac{b_{1,r,C}x}{n}\right) \leq \exp(-x) . \tag{6.87}$$

Then, notice that all conditions of Lemma 6.3 are satisfied, and that it gives by (6.68), for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|, A_-\right)$,

$$M_{1,r,C} \geq \left(1 - \frac{L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} > 0 \ . \tag{6.88}$$

In addition, observe that

$$\begin{aligned}
\sigma^2_{1,r,C} &\leq \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(\left(\psi_{1,M}\cdot(s_M - s)\right)^2\right) \\
&\leq 4\sup_{s\in\mathcal{F}_{(C,rC]}} \int (s_M - s)^2 f d\mu \\
&\leq 4\|f\| \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|(s_M - s)^2\right\| \qquad \text{by Cauchy-Schwarz inequality} \\
&\leq 4\|f\| \sup_{s\in\mathcal{F}_{(C,rC]}} \|s_M - s\|_\infty \times \sup_{s\in\mathcal{F}_{(C,rC]}} \|s_M - s\| \\
&\leq 4A_{3,M}\|f\| rC\sqrt{D} \qquad \text{by } (\mathbf{H2})
\end{aligned} \tag{6.89}$$

and

$$\begin{aligned}
b_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \|\psi_{1,M}\cdot(s_M - s) - P\psi_{1,M}\cdot(s_M - s)\|_\infty \\
&\leq 2\sup_{\|s\|_2\leq 1} \|\psi_{1,M}\cdot(s_M - s)\|_\infty \leq 4A_{3,M}\sqrt{rCD} \qquad \text{by } (\mathbf{H2}).
\end{aligned} \tag{6.90}$$

Hence, using (6.88), (6.89) and (6.90) in Inequality (6.87), we get for all $x > 0$ and all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|, A_-\right)$,

$$\mathbb{P}\left(\begin{array}{c} S_{1,r,C} \leq (1-\delta)\left(1 - \frac{L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \\ -\sqrt{\frac{8A_{3,M}\|f\|rC\sqrt{D}x}{n}} - \left(1+\frac{1}{\delta}\right)\frac{4A_{3,M}\sqrt{rCD}x}{n} \end{array}\right) \leq \exp(-x) \ .$$

Now, taking $x = \beta\ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}} \leq \frac{\sqrt{\ln n}}{D^{1/4}} \leq \frac{1}{2}$ for all $n \geq n_0\left(A_-\right)$ since $D \geq A_-\left(\ln n\right)^3$, and using $(\mathbf{H3})$, we can deduce by simple computations that a positive constant $L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|,\beta}$ exists such that, for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|, A_-\right)$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|,\beta}\times\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\right) \leq n^{-\beta} \ . \tag{6.91}$$

Finally, using (6.91) in (6.86) we get that, for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, \|f\|, A_-\right)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,\|f\|,\beta}\times\frac{\sqrt{\ln n}}{D^{1/4}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq n^{-\beta} \ ,$$

which concludes the proof. ■

In the case where $(\mathbf{H1})$ hold, we have the following result.

Lemma 6.8 *Let $r > 1$ and $C, \beta > 0$. Assume that **(H1)**, **(H2)** and **(H3)** hold. If positive constants A_-, A_+, A_l, A_u exist such that*

$$n \geq D \geq A_-\left(\ln n\right)^2 \quad \text{and} \quad A_l\frac{D}{n} \leq rC \leq A_u\frac{D}{n} \ ,$$

then a positive constant $L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty,\beta}$ exists such that, for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, A_\infty, A_-\right)$,

$$\mathbb{P}\left(\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty,\beta}\left(\sqrt{\frac{\ln n}{D}}\vee\frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq n^{-\beta} \ .$$

Proof. Start with

$$\begin{aligned}
&\sup_{s\in\mathcal{F}_{(C,rC]}} P_n\left(Ks_M - Ks\right)\\
&= \sup_{s\in\mathcal{F}_{(C,rC]}} \left\{\left(P_n - P\right)\left(Ks_M - Ks\right) + P\left(Ks_M - Ks\right)\right\}\\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(Ks - Ks_M\right)\\
&\geq \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right) - rC
\end{aligned} \tag{6.92}$$

and set

$$\begin{aligned}
S_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\\
M_{1,r,C} &= \mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)\right]\\
b_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|\psi_{1,M}\cdot\left(s_M - s\right) - P\psi_{1,M}\cdot\left(s_M - s\right)\right\|_\infty\\
\sigma^2_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \mathrm{Var}\left(\psi_{1,M}\cdot\left(s_M - s\right)\right).
\end{aligned}$$

By Klein-Rio's Inequality (7.50), we get, for all $\delta, x > 0$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1-\delta\right)M_{1,r,C} - \sqrt{\frac{2\sigma^2_{1,r,C}x}{n}} - \left(1+\frac{1}{\delta}\right)\frac{b_{1,r,C}x}{n}\right) \leq \exp\left(-x\right) . \tag{6.93}$$

Then, notice that all conditions of Lemma 6.3 are satisfied, and that it gives by (6.68), for all $n \geq n_0\left(A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, A_\infty, A_-\right)$,

$$M_{1,r,C} \geq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} > 0 . \tag{6.94}$$

In addition, observe that

$$\begin{aligned}
\sigma^2_{1,r,C} &\leq \sup_{s\in\mathcal{F}_{(C,rC]}} P\left(\psi_{1,M}\cdot\left(s_M - s\right)\right)^2\\
&\leq 4\sup_{s\in\mathcal{F}_{(C,rC]}} \int \left(s_M - s\right)^2 f d\mu\\
&\leq 4A_\infty \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|s_M - s\right\|^2 \qquad \text{by } (\mathbf{H1})\\
&\leq 4A_\infty rC
\end{aligned} \tag{6.95}$$

and

$$\begin{aligned}
b_{1,r,C} &= \sup_{s\in\mathcal{F}_{(C,rC]}} \left\|\psi_{1,M}\cdot\left(s_M - s\right) - P\psi_{1,M}\cdot\left(s_M - s\right)\right\|_\infty\\
&\leq 2\sup_{\|s\|_2\leq 1} \left\|\psi_{1,M}\cdot\left(s_M - s\right)\right\|_\infty \leq 4A_{3,M}\sqrt{rCD} \qquad \text{by } (\mathbf{H2}).
\end{aligned} \tag{6.96}$$

Hence, using (6.94), (6.95) and (6.96) in Inequality (6.93), we get for all $x > 0$ and for all $n \geq n_0 (A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, A_\infty, A_-)$,

$$\mathbb{P}\left(\begin{array}{c} S_{1,r,C} \leq (1-\delta)\left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \\ -\sqrt{\frac{8A_{3,M}\|f\|rC\sqrt{D}x}{n}} - \left(1+\frac{1}{\delta}\right)\frac{4A_{3,M}\sqrt{rCD}x}{n} \end{array} \right) \leq \exp(-x) \ .$$

Now, taking $x = \beta \ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}} \leq \sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}} \leq \frac{1}{2}$ for all $n \geq n_0 (A_-)$ since $D \geq A_- (\ln n)^2$, and using (**H3**), we deduce by simple computations that a positive constant $L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty,\beta}$ exists such that, for all $n \geq n_0 (A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, A_\infty, A_-)$,

$$\mathbb{P}\left(S_{1,r,C} \leq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty,\beta}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\right) \leq n^{-\beta} \ . \quad (6.97)$$

Finally, using (6.97) in (6.92) we get that, for all $n \geq n_0 (A_{3,M}, A_{\mathcal{K},-}, A_u, A_l, A_\infty, A_-)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{(C,rC]}} P_n (Ks_M - Ks) \leq \left(1 - L_{A_{3,M},A_{\mathcal{K},-},A_u,A_l,A_\infty,\beta}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} - rC\right) \leq n^{-\beta} \ ,$$

which concludes the proof. ■

Chapitre 7

Optimal excess risks bounds in regular contrast estimation

In this chapter, we derive upper and lower bounds with exact constants for the excess risk and its empirical counterpart on a fixed affine model of finite dimensional underlying vector space, in the general framework of regular contrast estimation exposed in Chapter 2. We refer to the introduction of Chapter 3 for detailed references on the subject to be addressed.
Section 7.1 is devoted to the framework of our study. We state our results in Section 7.2.3. We give some heuristics of the proofs in Section 7.3 and give the probabilistic tools needed, mainly concentration inequalities for the empirical process and other tools of the theory of probability in Banach spaces, in Section 7.4. The proofs of our results can be found at the end of this chapter.

7.1 Framework and notations

7.1.1 Regular contrast estimation

Let $(\mathcal{Z}, \mathcal{T})$ be a measurable space, P an unknown probability measure on $(\mathcal{Z}, \mathcal{T})$ and $\mathcal{S}$ a set of measurable functions from $(\mathcal{Z}, \mathcal{T})$ to $\mathbb{R}$. We also define $\xi_1, ..., \xi_n$ to be n independent random variables with common law P on $(\mathcal{Z}, \mathcal{T})$ and we take a generic random variable ξ of law P, independent of the sample $(\xi_1, ..., \xi_n)$. We consider a contrast K on $\mathcal{S}$ for the law P, that is a functional from $\mathcal{S}$ to $L_1^-(P)$,

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1^-(P) \\ s \longmapsto (Ks : z \longmapsto (Ks)(z)) \end{array} \right. ,$$

such that there exists a unique element $s_* \in \mathcal{S}$, called the target, satisfying

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) \quad \text{and} \quad P(Ks_*) < +\infty . \tag{7.1}$$

Let us recall the definition of the space $L_1^-(P)$ of real-valued measurable functions on $(\mathcal{Z}, \mathcal{T})$ whose negative part is of finite expectation with respect to P. The positive part of a real number $x \in \mathbb{R}$ is denoted $(x)_+ := \max\{x\ ,\ 0\} \geq 0$ and its negative part is $(x)_- := (-x)_+ = \max\{-x\ ,\ 0\} \geq 0$. We naturally extend these definitions to real-valued functions, and for a function f defined from $\mathcal{Z}$ to $\mathbb{R}$,

$$(f)_+ : z \in \mathcal{Z} \longmapsto (f(z))_+ \quad , \quad (f)_- : z \in \mathcal{Z} \longmapsto (f(z))_- \ .$$

Then, $L_1^-(P)$ is defined to be

$$L_1^-(P) = \left\{ f : \mathcal{Z} \to \mathbb{R} \ \ \mathcal{T}\text{-measurable}\ ;\ P(f)_- < +\infty \right\} .$$

Notice that expectation with respect to P is well-defined on $L_1^-(P)$, by writing for any $f \in L_1^-(P)$,

$$Pf := P(f)_+ - P(f)_- \in \overline{\mathbb{R}} ,$$

where $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. We also set $L_1(P)$ the set of integrable real-valued functions for the law P,

$$L_1(P) = \{f : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R} \; ; \; P|f| < +\infty\} ,$$

$L_2(P)$ is the set of square integrable real-valued functions for the law P,

$$L_2(P) = \left\{ f : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R} \; ; \; \|f\|_2 := \sqrt{P(f)^2} < +\infty \right\} ,$$

and $L_\infty(P)$ is the set real-valued functions essentially bounded on $\mathcal{Z}$ with respect to the law P,

$$L_\infty(P) := \{s : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R} \; ; \; \|s\|_\infty := \mathrm{essup}_{z \in \mathcal{Z}} (|s(z)|) < +\infty\} .$$

The target s_* is, according to (7.1), the minimizer of the risk $P(Ks)$ over the set $\mathcal{S}$. It is an unknown quantity as it depends on the law P. Our goal is to estimate the target s_* by using the sample $(\xi_1, ..., \xi_n)$. To that end, we consider a model $M \subset \mathcal{S} \bigcap L_\infty(P)$ and we take a M-estimator s_n on M, assumed to exist but non necessarily unique, satisfying

$$s_n \in \arg\min_{s \in M} P_n(Ks) \quad \text{with} \quad P_n(Ks_n) < +\infty \; a.s., \tag{7.2}$$

where

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$$

is the empirical distribution of the sample $(\xi_1, ..., \xi_n)$.
Moreover, we assume that the contrast K is regular for the model M and the law P. A precise definition of a regular contrast can be found in Section 2.2 of Chapter 2, that recall now. First, we assume that there exists a unique projection s_M of s_* on M, defined to be

$$s_M = \arg\min_{s \in M} P(Ks) \quad \text{with} \quad P(Ks_M) < +\infty . \tag{7.3}$$

In addition, for all $s \in M$ and P-almost all $z \in \mathcal{Z}$, the following expansion hold,

$$Ks(z) - Ks_M(z) = \psi_0^s + \psi_{1,M}(z)(s - s_M)(z) + \psi_2(\psi_{3,M}(z)(s - s_M)(z)) , \tag{7.4}$$

where ψ_0^s is a constant depending on s but not on z, $\psi_{1,M}$ and $\psi_{3,M}$ are functions defined on $\mathcal{Z}$ not depending on s and not identically equal to 0 satisfying $\psi_{1,M} \in L_2(P)$, $\psi_{3,M} \in L_\infty(P)$ and ψ_2 is a function not depending on s, defined on a subset $\mathcal{D}_2 \subseteq \mathbb{R}$ such that $0 \in \mathring{\mathcal{D}}_2$, where $\mathring{\mathcal{D}}_2$ denotes the interior of $\mathcal{D}_2$, $\psi_2(\mathcal{D}_2) \subseteq \overline{\mathbb{R}}$ and $\psi_2(0) = 0$. Moreover, there exists $L_2 > 0$ such that for all $\delta \in \left[0, L_2^{-1}\right]$, it holds $[-\delta, \delta] \subset \mathcal{D}_2$ and for all $(x, y) \in [-\delta, \delta]$,

$$|\psi_2(x) - \psi_2(y)| \leq L_2 \delta |x - y| . \tag{7.5}$$

Thirdly, if we denote

$$\widetilde{M_0} := \mathrm{Span}\{s - s_M \; ; \; s \in M\} ,$$

then there exists an Hilbertian norm $\|\cdot\|_{H,M}$ on $\widetilde{M_0}$ and positive constants $A_H, L_H > 0$ such that

$$\|\cdot\|_2 \leq A_H \|\cdot\|_{H,M} \tag{7.6}$$

and for all $\delta \in \left[0, L_H^{-1}\right]$, for all $s \in M$ such that $\|s - s_M\|_\infty \leq \delta \leq L_H^{-1}$, it holds

$$(1 - L_H \delta) \|s - s_M\|_{H,M}^2 \leq P(Ks - Ks_M) \leq (1 + L_H \delta) \|s - s_M\|_{H,M}^2 . \tag{7.7}$$

We further assume in this chapter that the model M is affine, that is

$$M_0 := \{s - s_M \ ; \ s \in M\} \tag{7.8}$$

is a linear vector space, and we demand that M_0 has a finite linear dimension that we denote D. This gives $M_0 = \widetilde{M_0}$ and so, $\|\cdot\|_{H,M}$ is an Hilbertian norm on M_0. By abuse, M_0 is also called a model.

The framework described above contains least-squares regression and least-squares estimation of density on finite dimensional linear models, the latter being indeed special cases of affine spaces, and maximum likelihood estimation of density on histograms, as further explained in Chapter 2. Considering the case of a model M of histograms for maximum likelihood estimation, see Chapter 5, one has to restrict to histogram densities with respect to some known probability measure μ on $(\mathcal{Z}, \mathcal{T})$,

$$M = \left\{ s = \sum_{I \in \Lambda_M} \beta_I \mathbf{1}_I \ ; \ (\beta_I)_{I \in \Lambda_M} \in \mathbb{R}^\Lambda \ , \ s \geq 0 \ , \ \int_{\mathcal{Z}} s d\mu = 1 \right\} \ ,$$

where Λ_M is some finite partition on $\mathcal{Z}$ and $\Lambda = \mathrm{Card}\,(\Lambda_M)$. Thus, for any $s \in M$, it holds

$$\int_{\mathcal{Z}} (s - s_M)\, d\mu = 0 \ , \tag{7.9}$$

a property also satisfied on $\widetilde{M_0} = \mathrm{Span}\,\{s - s_M \ ; \ s \in M\}$. However, the set

$$M_0 = \{s - s_M \ ; \ s \in M\}$$

is not a linear vector space, since for any $s \in M$, $s \geq 0$. In fact, it is only "star-shaped" at 0. But in Chapter 5 we further assume that there exists a positive constant $A_{\min}$ such that

$$\inf_{z \in \mathcal{Z}} s_M(z) \geq A_{\min} > 0 \ .$$

This allows to recover some "linearity" on the set M_0 locally around the projection s_M, in the sense that it holds

$$\widetilde{M_0} \bigcap B_{(M,L_\infty)}(s_M, A_{\min}) = M_0 \bigcap B_{(M,L_\infty)}(s_M, A_{\min}) \ ,$$

where $B_{(M,L_\infty)}(s_M, A_{\min}) := \{s \in M \ ; \ \|s - s_M\|_\infty \leq A_{\min}\}$. As everything happens in the subset $B_{(M,L_\infty)}(s_M, A_{\min})$ of M due to the assumption of consistency in sup-norm of the maximum likelihood estimator towards the projection s_M made in Chapter 5 and also in this chapter, see Section 7.1.3, straightforward adaptations of the formalism developed in the present chapter allow to consider the maximum likelihood estimation of density on histograms as a special case of our general framework.

7.1.2 Excess risks

Our aim is to study the performance of the M-estimator s_n in terms of excess risk, defined to be the following random quantity

$$P(Ks_n - Ks_*) \geq 0 \ . \tag{7.10}$$

Notice that the excess risk given in (7.10) is well-defined. Indeed, by (7.1), we have $P(Ks_*) < +\infty$, and as $s_* \in \mathcal{S}$ we also have $P(Ks_*)_- < +\infty$, and thus $|P(Ks_*)| \leq P(|Ks_*|) < +\infty$, or in other words $Ks_* \in L_1(P)$. Moreover, as $s_n \in M \subset \mathcal{S}$ it holds $P(Ks_n) \in \overline{\mathbb{R}}$ and we conclude that

$$P(Ks_n - Ks_*) = P(Ks_n) - P(Ks_*) \in \overline{\mathbb{R}} \ .$$

The fact that the excess risk is nonnegative directly follows from the definition of the target s_*, namely s_* is the minimizer of the risk over $\mathcal{S}$.
The excess risk of the M-estimator s_n decomposes into the sum of the bias of the model and the excess risk of the estimator on M :

$$P\left(Ks_n - Ks_*\right) = P\left(Ks_n - Ks_M\right) + P\left(Ks_M - Ks_*\right) \ .$$

The excess risk on M, $P\left(Ks_n - Ks_M\right)$ is well-defined by the same arguments showing that the excess risk $P\left(Ks_n - Ks_*\right)$ is well-defined. As the projection s_M is the minimizer of the risk over the model M, it holds

$$P\left(Ks_n - Ks_M\right) \geq 0 \ .$$

Furthermore, since $Ks_*, Ks_M \in L_1\left(P\right)$, the bias of the model is finite, $P\left(Ks_M - Ks_*\right) < +\infty$ and as s_* is the minimizer of the risk over $\mathcal{S}$, we have

$$P\left(Ks_M - Ks_*\right) \geq 0 \ .$$

The bias term is deterministic and we only focus here on the excess risk on M of the estimator s_n. We do not discuss on the possible behaviors of the bias of the model, neither on the trade-off that can be achieved between the bias and the excess risk on M.
Another key quantity will be studied in this chapter, which is closely related to the excess risk on M of the M-estimator, namely the empirical excess risk on M of the M-estimator, defined to be

$$P_n\left(Ks_M - Ks_n\right) \geq 0 \ .$$

We notice that since $Ks_n \in L_1^-\left(P\right)$ and $P_n\left(Ks_n\right) < +\infty \ a.s.$ by definition of s_n given in (8.2), we have $\left|P_n\left(Ks_n\right)\right| < +\infty \ a.s.$. As $Ks_M \in L_1\left(P\right)$, we have $\left|P_n\left(Ks_M\right)\right| < +\infty \ a.s.$ and so the empirical excess risk on the estimator is well-defined. The fact that this quantity is nonnegative directly follows from the definition of s_n, as s_n is the minimizer of the empirical risk over M.

7.1.3 Assumptions on the model

We state here the two assumptions needed on the model M in order to derive the results of Section 7.1.2. These assumptions, that relate the $L_2\left(P\right)$-structure and the $L_\infty\left(P\right)$-structure of linear models, were first formulated by Birgé and Massart in [25] - see also Barron, Birgé and Massart [13] and Massart [61] Section 7.4.2 - in order to derive accurate excess risk bounds in a general M-estimation setting. The authors also generalized these assumptions to non-linear cases by considering suitable entropy numbers, see [25].
Our first assumption requires the following preliminary definition. Recall that the model M is affine, with underlying vector space M_0 pointed at the projection s_M and that $\left\|\cdot\right\|_{H,M}$ is an Hilbertian norm on M_0. Moreover, M_0 has a finite linear dimension D.

Definition 7.1 *The **unit envelope** Ψ_M of the model M, for the Hilbertian norm $\left\|\cdot\right\|_{H,M}$, is a function defined on $\mathcal{Z}$, such that for any $z \in \mathcal{Z}$,*

$$\Psi_M\left(z\right) = \frac{1}{\sqrt{D}} \sup_{t \in M_0, \left\|t\right\|_{H,M} \leq 1} \left|t\left(z\right)\right| \ . \tag{7.11}$$

Since M_0 is a finite dimensional real vector space, the supremum in (7.11) can also be taken over a countable subset of M_0, so Ψ_M is a measurable function.

We must require in the following assumption that the unit envelope of the model M_0 defined in (8.8) is uniformly bounded on $\mathcal{Z}$.

- (**A1**) The unit envelope of M_0 is uniformly bounded on $\mathcal{Z}$: a positive constant A_Ψ exists such that

$$\|\Psi_M\|_\infty \leq A_\Psi < \infty \ .$$

The following assumption is stronger than (**A1**).

- (**A2**) Existence of a **localized basis** in $\left(M_0, \|\cdot\|_{H,M}\right)$: there exists an orthonormal basis $\varphi = (\varphi_k)_{k=1}^D$ in $\left(M_0, \|\cdot\|_{H,M}\right)$ that satisfies, for a positive constant $r_M(\varphi)$ and all $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$,

$$\left\| \sum_{k=1}^D \beta_k \varphi_k \right\|_\infty \leq r_M(\varphi) \sqrt{D} \, |\beta|_\infty \ ,$$

where $|\beta|_\infty = \max\{|\beta_k| \, ; k \in \{1, ..., D\}\}$ is the sup-norm of the D-dimensional vector β.

Remark 7.1 *(**A2**) implies (**A1**) and if (**A1**) holds then $A_\Psi = r_M(\varphi)$ is convenient.*

In Birgé and Massart [25] and also in Section 7.4.2 of Massart [61], it is shown that models of histograms, piecewise polynomials and compactly supported wavelets are typical examples of models with localized basis for the L_2 (Leb) structure, considering for instance that $\mathcal{Z} \subset \mathbb{R}^k$. Moreover, Fourier expansions on the unit interval with respect to the Lebesgue measure Leb satisfy (**A1**) but their index of localization $r_M(\varphi)$ are typically of order $L\sqrt{D}$ for some positive constant L.
In Chapter 3, where we study the least-squares regression on a fixed linear model, we show that histogram models are endowed with a localized basis structure in $L_2\left(P^X\right)$, where P^X is the marginal law of the explicative variable X, under the assumption that the finite partition defining the model is lower-regular with respect to the law P^X, see Section 3.4.1 of Chapter 3. We also consider the case of piecewise polynomials defined on a finite partition of the unit interval and that are uniformly bounded in their degree. We show in Section 3.5.1 of Chapter 3 that if the law P^X has a density, with respect to the Lebesgue measure on the unit interval, which is uniformly bounded away from zero and if the partition is lower-regular with respect to the Lebesgue measure then the piecewise polynomials are again endowed with a localized basis structure in $L_2\left(P^X\right)$.

7.2 Excess risks bounds

We state in Section 7.2.3 below the main results of this chapter. We show in particular that when the considered model has a "reasonable" dimension, the excess risk on M of the M-estimator is equivalent to the empirical excess risk on M, which is a keystone to prove the slope heuristics of Birgé and Massart. We find rates of convergence that are optimal at the first order. These rates involve a key quantity that relates the structure of the image by the regular contrast K of the model M with the unknown law P and that we call the complexity of the model, see Section 7.2.2. To derive results of Section 7.2.3, we need to assume that the M-estimator is consistent in sup-norm towards the projection of the target onto the model. This assumption is stated in Section 7.2.1.

7.2.1 Assumption of consistency in sup-norm

The following assumption states that the M-estimator s_n is consistent towards the projection s_M of the target onto the model, at a rate not slower than $(\ln n)^{-1/2}$.

- (**A3**) Assumption of consistency in sup-norm **:** for any $A_+ > 0$, if M_0 is a model of dimension D satisfying

$$D \leq A_+ \frac{n}{(\ln n)^2} \ ,$$

then for all $\alpha > 0$, we can find a positive integer n_1 and a positive constant A_{cons} satisfying the following property : there exists $R_{n,D,\alpha} > 0$ depending on D, n and α, such that

$$R_{n,D,\alpha} \leq \frac{A_{cons}}{\sqrt{\ln n}} \tag{7.12}$$

and by setting

$$\Omega_{\infty,\alpha} = \left\{ \|s_n - s_M\|_\infty \leq R_{n,D,\alpha} \right\} \ , \tag{7.13}$$

it holds, for all $n \geq n_1$,

$$\mathbb{P}\left[\Omega_{\infty,\alpha}\right] \geq 1 - n^{-\alpha} \ . \tag{7.14}$$

In Chapter 3, we prove in a regression framework that the least-squares estimator achieves assumption (**A3**) - which is denoted by (**H5**) in Chapter 3 - for suitable histogram models and models of piecewise polynomials, see Sections 3.4.2 and 3.5.2 respectively. In such cases, we need to assume, among other things, that the partitions defining the models are lower-regular with respect to the law P^X of the explicative variable X. In Chapter 5, we show that the maximum likelihood estimator on histograms is consistent towards the Kullback-Leibler projection of the target onto the model, when the partitions defining the models are lower-regular with respect to a measure of reference μ, see Section 5.3.1 of Chapter 5. We also notice that in the case of least-squares estimation of density, studied in Chapter 6, the consistency assumption (**A3**) is not needed, which is due to the fact that the least-squares density contrast is linear - which means that $\psi_2 \equiv 0$ where ψ_2 is defined in (7.4).

7.2.2 Complexity of the model

Recall that the model M is assumed to be affine, with underlying vector space M_0 pointed at the projection s_M. Moreover, M_0 has a finite linear dimension D and there exists an Hilbertian norm $\|\cdot\|_{H,M}$ on M_0, such that for some positive constant A_H and any $t \in M_0$,

$$\|t\|_2 \leq A_H \|t\|_{H,M} \ .$$

The unit envelope of the model M for the norm $\|\cdot\|_{H,M}$ is written Ψ_M, see Defintion 7.1.

Definition 7.2 *The **complexity** of the model M under the regular contrast K and the law P is written $\mathcal{C}_M$ and is defined by*

$$\mathcal{C}_M = \frac{1}{4} D \times \mathcal{K}_{1,M}^2 \ ,$$

where

$$\mathcal{K}_{1,M}^2 := \frac{1}{D} \left[P\left(\psi_{1,M}^2 \cdot \sup_{t \in M_0,\ \|t\|_{H,M} \leq 1} t^2 \right) - \sup_{t \in M_0,\ \|t\|_{H,M} \leq 1} \left[P\left(\psi_{1,M} \cdot t\right) \right]^2 \right] \tag{7.15}$$

$$= P\left(\psi_{1,M}^2 \cdot \Psi_M^2\right) - \frac{1}{D} \sup_{t \in M_0,\ \|t\|_{H,M} \leq 1} \left[P\left(\psi_{1,M} \cdot t\right) \right]^2 \geq 0 \ . \tag{7.16}$$

The quantity $\mathcal{K}_{1,M} = \sqrt{\mathcal{K}_{1,M}^2} \geq 0$ is called the normalized complexity of the model M.

We shall prove that the quantity given in (7.16) is well-defined and nonnegative. Let us take an orthonormal basis $(\varphi_k)_{k=1}^D$ in $\left(M_0, \|\cdot\|_{H,M}\right)$. By using Cauchy-Schwarz inequality in (7.11), we have

$$\Psi_M = \sqrt{\frac{1}{D}\sum_{k=1}^{D} \varphi_k^2} \ . \tag{7.17}$$

Now, as $M \subset L_\infty(P)$ we have $M_0 \subset L_\infty(P)$ and by (7.17) it moreover holds $\Psi_M \subset L_\infty(P)$. Since $\psi_{1,M} \in L_2(P)$, we thus get

$$P\left(\psi_{1,M}^2 \cdot \Psi_M^2\right) < +\infty \ . \tag{7.18}$$

Furthermore, since $M_0 \subset L_2(P)$ and $\psi_{1,M} \in L_2(P)$, $P(\psi_{1,M} \cdot t)$ is well-defined for any $t \in M_0$. Moreover,

$$\begin{aligned} \sup_{t \in M_0,\ \|t\|_{H,M} \leq 1} \left[P(\psi_{1,M} \cdot t)\right]^2 &= \sup_{(\beta_k)_{k=1}^D \in \mathbb{R}^D,\ \sum \beta_k^2 \leq 1} \left[\sum_{k=1}^{D} \beta_k P(\psi_{1,M} \cdot \varphi_k)\right]^2 \\ &= \sum_{k=1}^{D} \left[P(\psi_{1,M} \cdot \varphi_k)\right]^2 \ , \end{aligned} \tag{7.19}$$

where the last equality follows from Cauchy-Schwarz inequality. From (7.19), we deduce that $\sup_{t \in M_0,\ \|t\|_{H,M} \leq 1} \left[P(\psi_{1,M} \cdot t)\right]^2 < +\infty$ and so, by (7.16), we get $\mathcal{K}_{1,M}^2 < +\infty$. By combining (7.16), (7.17) and (7.19) we also deduce that

$$\mathcal{K}_{1,M}^2 = \frac{1}{D}\sum_{k=1}^{D} \mathrm{Var}(\psi_{1,M} \cdot \varphi_k) \geq 0 \ , \tag{7.20}$$

which also proves that $\mathcal{K}_{1,M}^2$ is nonnegative.

Remark 7.2 *The model M is affine and thus star-shaped in the sense of Definition 2.9 of chapter 2. By Proposition 2.2 of Chapter 2, for each $k \in \{1, ..., D\}$, there exists a constant $a_k \in \mathbb{R}$ such that $\psi_{1,M} \cdot \varphi_k$ is defined up to the constant a_k on $\mathcal{Z}$. Now, since $\mathrm{Var}(\psi_{1,M} \cdot \varphi_k)$ is independent of the value of a_k, we deduce that the normalized complexity $\mathcal{K}_{1,M}$ is independent of the choice of $\psi_{1,M}$ whenever M is affine, and so does the complexity $\mathcal{C}_M$.*

Let us now assume that $\psi_{1,M} \in L_\infty(P)$ and $\psi_{3,M} \in L_\infty(P)$.

- (**A4**) Coefficients $\psi_{1,M}$ and $\psi_{3,M}$ appearing in the expansion of contrast are uniformly bounded on $\mathcal{Z}$: There exists a positive constant A_1 such that

$$\|\psi_{1,M}\|_\infty \leq A_1 < +\infty \ . \tag{7.21}$$

$$0 < A_3 = \|\psi_{3,M}\|_\infty < +\infty \ . \tag{7.22}$$

Since we have by (7.6) and (7.22), for any $k \in \{1, ..., D\}$,

$$\mathrm{Var}(\psi_{1,M} \cdot \varphi_k) \leq P\left(\psi_{1,M}^2 \cdot \varphi_k^2\right) \leq A_1^2 \|\varphi_k\|_2^2 \leq (A_1 A_H)^2 \|\varphi_k\|_{H,M}^2 = (A_1 A_H)^2 \ ,$$

we get

$$\mathcal{K}_{1,M} \leq A_1 A_H \ . \tag{7.23}$$

We further need the following assumption, ensuring that the normalized complexity $\mathcal{K}_{1,M}$ indeed behaves like a constant.

- (**A5**) Lower bound on the normalized complexity : a positive constant $A_{\mathcal{K}}$ exists such that

$$\mathcal{K}_{1,M} \geq A_{\mathcal{K}} > 0 \ . \tag{7.24}$$

In the regression framework, assumption (**A4**) is satisfied when the response variable Y is almost surely bounded, see Section 3.3.3 of Chapter 3. In fact, (7.22) is in this case automatically satisfied as we fix $\psi_{3,M} = 1$. In Chapter 5, considering maximum likelihood estimation of density, we have

$$\psi_{1,M} = \psi_{3,M} = \frac{1}{s_M} \ ,$$

and considering a linear model made of histograms, we show that (**A4**) is satisfied if the density s_* to be estimated is uniformly bounded on $\mathcal{Z}$, see Section 5.5.3. In the case of least-squares estimation of density, assumption (A4) is automatically satisfied since $\psi_{1,M}$ is a constant function and $\psi_{3,M}$ is equal to zero.
We show in Section 3.3.3 of Chapter 3 that assumption (**A5**) is satisfied in the regression setting if the noise level or the unit envelope are uniformly bounded away from zero. In our general regular contrast setting, if we assume that there exists $A_{1,-}$, $A_{H,-} > 0$ and $c \in (0,1)$ such that

$$\inf_{z \in \mathcal{Z}} |\psi_{1,M}| \geq A_{1,-} > 0 \ , \tag{7.25}$$

$$\|\cdot\|_2 \geq A_{H,-} \|\cdot\|_{H,M} \tag{7.26}$$

and for some orthonormal basis $(\varphi_k)_{k=1}^D$ of $\left(M, \|\cdot\|_{H,M}\right)$ it holds

$$\sup_{k \in \{1,\dots,D\}} \mathbb{E}\left[|\varphi_k|\right] \leq c \frac{A_{1,-} A_{H,-}}{A_1} \ ,$$

then it is easy to check that (**A5**) holds with $A_{\mathcal{K}} = \left(1 - c^2\right) \left(A_{1,-} A_{H,-}\right)^2 > 0$.

7.2.3 Theorems

We state now the main results of this chapter. The following theorem generalizes Theorem 3.1 of Chapter 3, from least-squares regression to regular contrast estimation.

Theorem 7.1 *Let $A_+, A_-, \alpha > 0$ and let $(K, \mathcal{S}, P)$ be such that $K : \mathcal{S} \to L_1^-(P)$ is a contrast on $\mathcal{S}$ for the law P. Take a model $M \subset \mathcal{S} \bigcap L_\infty(P)$ and assume that K is regular on M for the law P. Assume moreover that M is an affine space, with underlying linear vector space M_0 pointed at the projection s_M. Assume that M_0 has a finite linear dimension D and that **(A2)**, **(A3)**, **(A4)** and **(A5)** hold. Take $\varphi = (\varphi_k)_{k=1}^D$ an orthonormal basis of $\left(M_0, \|\cdot\|_{H,M}\right)$ satisfying **(A2)**. If it holds*

$$A_- (\ln n)^2 \leq D \leq A_+ \frac{n}{(\ln n)^2} \ ,$$

*then a positive finite constant A_0 exists, only depending on α, A_-, on the constants L_2, A_H, L_H defined in Section 7.1.1 and on the constants $r_M(\varphi), A_1, A_{\mathcal{K}}$ respectively defined in the assumptions **(A2)**, **(A4)** and **(A5)**, such that by setting*

$$\varepsilon_n = A_0 \max\left\{ \left(\frac{\ln n}{D}\right)^{1/4}, \ \left(\frac{D \ln n}{n}\right)^{1/4}, \ \sqrt{R_{n,D,\alpha}} \right\} \ , \tag{7.27}$$

we have for all $n \geq n_0 \left(A_-, A_+, L_2, A_H, L_H, r_M(\varphi), A_1, A_{\mathcal{K}}, A_3, A_{cons}, n_1, \alpha\right)$,

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \geq \left(1 - \varepsilon_n\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2\right] \geq 1 - 5n^{-\alpha} \ , \tag{7.28}$$

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \leq \left(1 + \varepsilon_n\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2\right] \geq 1 - 5n^{-\alpha} \ , \tag{7.29}$$

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\right) \geq \left(1 - \varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2\right] \geq 1 - 2n^{-\alpha} \ , \tag{7.30}$$

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\right) \leq \left(1 + \varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2\right] \geq 1 - 3n^{-\alpha} \ . \tag{7.31}$$

In addition, when ***(A3)*** *does not hold, but* ***(A2)***, ***(A3)***, ***(A4)*** *and* ***(A5)*** *hold, we still have for all* $n \geq n_0 \left(A_-, A_+, L_2, A_H, L_H, r_M(\varphi), A_1, A_{\mathcal{K}}, A_3, \alpha\right)$,,

$$\mathbb{P}\left(P_n\left(Ks_M - Ks_n\right) \geq \left(1 - A_0 \max\left\{\sqrt{\frac{\ln n}{D}}, \sqrt{\frac{D \ln n}{n}}\right\}\right) \frac{D}{4n} \mathcal{K}_{1,M}^2\right) \geq 1 - 2n^{-\alpha} \ . \tag{7.32}$$

Theorem 7.2 *Let* $A_-, \alpha > 0$ *and let* $(K, \mathcal{S}, P)$ *be such that* $K : \mathcal{S} \to L_1^-(P)$ *is a contrast on* $\mathcal{S}$ *for the law* P. *Take a model* $M \subset \mathcal{S} \bigcap L_\infty(P)$ *and assume that* K *is regular on* M *for the law* P. *Assume moreover that* M *is an affine space, with underlying linear vector space* M_0 *pointed at the projection* s_M. *Assume that* M_0 *has a finite linear dimension* D *and that* ***(A1)***, ***(A3)*** *and* ***(A4)*** *hold. If it holds*

$$1 \leq D \leq A_+ \frac{n}{(\ln n)^2} \ ,$$

then a positive finite constant A_u *exists, only depending on* α, A_-, *on the constants* L_2, A_H, L_H *defined in Section 7.1.1 and on the constants* $A_\Psi, A_1, A_{\mathcal{K}}$ *respectively defined in the assumptions* ***(A2)***, ***(A4)*** *and* ***(A5)***, *such that for all* $n \geq n_0 \left(A_{cons}, n_1\right)$,

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \tag{7.33}$$

and

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\right) \geq A_u \frac{D \vee \ln n}{n}\right] \leq 3n^{-\alpha} \ . \tag{7.34}$$

7.3 The problem of upper and lower bounds

7.3.1 Rewriting the problem

We reformulate here the mathematical problems of upper and lower bounds in order to achieve a formulation involving tractable mathematical quantities. We shall emphasize the fact that no references to a linear structure of the model are needed, so that this is still valid for other non-linear M-estimation problems such as supervised classification or level set estimation. This allows us also to identify the keystones of the proofs in order to propose directions of research to generalize our results.

The question of upper bound for the true excess risk on the model can be stated as follows. Find the smallest level C such that the probability

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \geq C\right]$$

is bounded from above by a negative power of n. Traditionally, the proofs begin by symmetrizing the true excess risk with respect to the empirical one. Indeed, since by definition of s_n, it holds $P_n(Ks_n - Ks_M) \leq 0$, we have

$$P(Ks_n - Ks_M) \leq (P - P_n)(Ks_n - Ks_M) \tag{7.35}$$

so that

$$\begin{aligned}&\mathbb{P}[P(Ks_n - Ks_M) \geq C]\\ &\leq \mathbb{P}[(P - P_n)(Ks_n - Ks_M) \geq C].\end{aligned}$$

Then, one can use various empirical process techniques. For example, by observing that

$$(P - P_n)(Ks_n - Ks_M) \leq \sup_{s \in \mathcal{G}} |(P - P_n)(Ks - Ks_M)| \tag{7.36}$$

where $\mathcal{G}$ is a subset of M, for instance a subset of uniformly bounded functions, we can apply chaining techniques to evaluate the mean of the supremum of the empirical process and Talagrand's type concentration inequalities to control the deviations from the mean. But there is generally a huge loss in (7.36). In fact, if we have an explicit bound $\phi_n(\delta)$ such that, with high probability, it holds

$$\sup_{\{s, P(Ks - Ks_M) \leq \delta\}} |(P - P_n)(Ks - Ks_M)| \leq \phi_n(\delta) \tag{7.37}$$

then, using (7.35) we get

$$P(Ks_n - Ks_M) \leq \phi_n(P(Ks_n - Ks_M))$$

so with the same probability as for (7.37), $P(Ks_n - Ks_M)$ is bounded from above by the largest solution of $\delta \leq \phi_n(\delta)$.
Now, bounding the supremum of the empirical process requires to evaluate the variance term of the indexes, so that inequality (7.37) asks for a control of the variance at a point of the model by the excess risk at the same point. This kind of relation is called a margin relation, the name coming from the classification setting, see [75]. It is usually of the form

$$\mathrm{Var}(Ks - Ks_M) \leq \kappa (P(Ks - Ks_M))^{\alpha}$$

for some $\kappa > 0$ and $0 < \alpha \leq 1$.
This refinement of the control of local increments leads to sharper and faster rates of convergence, that are often minimax, with sometimes discussions on some extra log factor, see [39] and [62]. But the use of (7.35) has two major drawbacks with respect to our problem. First, our aim is to upper and lower bound the true and the empirical excess risks, non only at the right and same rate, but with the right constant in front of the speed in order to prove that

$$P(Ks_n - Ks_M) \sim P_n(Ks_M - Ks_n) \ .$$

Unfortunately, the use of (7.35) would multiply the constant at least by two. Another drawback is that the control of the excess risk in (7.35) is useless for lower bounds where the goal is to find the largest C such that the probability

$$\mathbb{P}[P(Ks_n - Ks_M) \leq C]$$

is bounded from below by a negative power of n.

We recall that by definition

$$\begin{aligned} s_n &\in \arg\min_{s\in M} P_n Ks \\ &= \arg\min_{s\in M} P_n\left(Ks - Ks_M\right) \\ &= \arg\min_{s\in M} \left\{\left(P_n - P\right)\left(Ks - Ks_M\right) + P\left(Ks - Ks_M\right)\right\} \ . \end{aligned}$$

So we are dealing with the argument of the minimum over a class of functions of an empirical process drifted, the deterministic drift being the excess risk. Moreover, we want to localize this argument by the value of his drift. So a natural idea, that can be found in [44], [39] or in the pealing lemma of [62], is to cut the class of functions into slices of drift, in order to estimate the values of the empirical process on each slice as sharply as possible and compare these values to determine their argument of minimum. In particular this requires to obtain upper and lower bounds for each slice, on the contrary of (7.37) where one just needs upper bounds. This is in essence what we do in our proofs. More precisely, if we set

$$\begin{aligned} M_C &= \left\{s \in M; P\left(Ks - Ks_M\right) \leq C\right\} \\ M_{>C} &= \left\{s \in M; P\left(Ks - Ks_M\right) > C\right\} \\ &= M_C^c \end{aligned}$$

and, for an interval I,

$$M_I = \left\{s \in M; P\left(Ks - Ks_M\right) \in I\right\}$$

then we can write, assuming that s_n exists,

$$\begin{aligned} &\left\{P\left(Ks_n - Ks_M\right) \leq C\right\} \\ &\leq \left\{\inf_{s\in M_C} P_n\left(Ks - Ks_M\right) \leq \inf_{s\in M_{>C}} P_n\left(Ks - Ks_M\right)\right\} \\ &= \left\{\begin{array}{l} \sup_{s\in M_C}\left\{\left(P - P_n\right)\left(Ks - Ks_M\right) - P\left(Ks - Ks_M\right)\right\} \\ \geq \sup_{s\in M_{>C}}\left\{\left(P - P_n\right)\left(Ks - Ks_M\right) - P\left(Ks - Ks_M\right)\right\} \end{array}\right\} \end{aligned} \tag{7.38}$$

and also

$$\begin{aligned} &\left\{P\left(Ks_n - Ks_M\right) \geq C\right\} \\ &\leq \left\{\inf_{s\in M_C} P_n\left(Ks - Ks_M\right) \geq \inf_{s\in M_{>C}} P_n\left(Ks - Ks_M\right)\right\} \\ &= \left\{\begin{array}{l} \sup_{s\in M_C}\left\{\left(P - P_n\right)\left(Ks - Ks_M\right) - P\left(Ks - Ks_M\right)\right\} \\ \leq \sup_{s\in M_{>C}}\left\{\left(P - P_n\right)\left(Ks - Ks_M\right) - P\left(Ks - Ks_M\right)\right\} \end{array}\right\} \ . \end{aligned} \tag{7.39}$$

Then we intend to apply empirical processes techniques in order to handle these suprema. Notice that the formulation (7.38) and (7.39) is rather general and free from any reference to the structure of the contrast or the structure of the model.
Moreover, after working out this approach, we are eventually able to explain the accurate relation that can occur between the true excess risk and the empirical one. Indeed, if one can find upper and lower bounds for the true excess risk, that is C_+, $C_- > 0$ such that, with high probability,

$$C_+ \geq P\left(Ks_n - Ks_M\right) \geq C_- \ ,$$

then by (7.38) and (7.39), we have, with the same probability,

$$\sup_{s\in M} P_n\left(Ks_M - Ks\right) = \sup_{s\in M_{[C_-,C_+]}} P_n\left(Ks_M - Ks\right) \ . \tag{7.40}$$

Since we have, by definition of s_n,

$$\sup_{s\in M} P_n\left(Ks_M - Ks\right) = P_n\left(Ks_M - Ks_n\right)$$

we deduce that, if we are able to upper and lower bound with high probability the right-hand side of (7.40), where the infimum is taken over a localized slice of the model, then we can derive upper and lower bounds for the empirical excess risk. In addition, if $C_+ \leq (1+\varepsilon)\, C_*$ and $C_- \geq (1-\varepsilon)\, C_*$ for some very small $\varepsilon > 0$ and

$$\sup_{s\in M_{[C_-,C_+]}} P_n\left(Ks_M - Ks\right) \approx C_* \tag{7.41}$$

with high probability, then we have $P\left(Ks_n - Ks_M\right) \approx P_n\left(Ks_M - Ks_n\right)$. We recognize in (7.41) the classical argument of *fixed point* used to derive upper bounds of the excess risk in the context of margin relations, see for example [62] or [44].
We now turn in the following section to a more precise heuristics in the case of bounded regression with a linear model.

7.3.2 Heuristics of the proofs

As we have seen in Section 7.3 above, we have

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \leq C\right] \leq \mathbb{P}\left[\sup_{s\in M_C} P_n\left(Ks_M - Ks\right) \geq \sup_{s\in M_{>C}} P_n\left(Ks_M - Ks\right)\right] .$$

Since we have, for any $r > 1$,

$$\sup_{s\in M_{>C}} P_n\left(Ks_M - Ks\right) \geq \sup_{s\in M_{(C,rC]}} P_n\left(Ks_M - Ks\right) ,$$

we deduce that

$$\mathbb{P}\left[P\left(Ks_n - Ks_M\right) \leq C\right] \leq \mathbb{P}\left[\sup_{s\in M_C} P_n\left(Ks_M - Ks\right) \geq \sup_{s\in M_{(C,rC]}} P_n\left(Ks_M - Ks\right)\right] .$$

Now, by setting

$$S_L = M_{[L,L]} = \left\{s \in M, P\left(Ks - Ks_M\right) = L\right\}$$

we have

$$\sup_{s\in M_C} P_n\left(Ks_M - Ks\right) = \sup_{0\leq L\leq C} \sup_{s\in S_L} P_n\left(Ks_M - Ks\right)$$

and

$$\sup_{s\in M_{]C,rC]}} P_n\left(Ks_M - Ks\right) = \sup_{C<L\leq rC} \sup_{s\in S_L} P_n\left(Ks_M - Ks\right) .$$

As a consequence, the core of the proof is to understand how fast grows $\sup_{s\in S_L} P_n\left(Ks_M - Ks\right)$, and the lower bound as well as the upper bound will be very close to

$$C_* = \arg\max_L \sup_{s\in S_L} P_n\left(Ks_M - Ks\right) .$$

To fix ideas, we take the example of least-squares regression. We recall that in this case,

$$Ks_M - Ks = \psi_{1,M} \cdot \left(s_M - s\right) - \left(s - s_M\right)^2$$

and

$$P\left(Ks - Ks_M\right) = P\left(s - s_M\right)^2 = \left\|s - s_M\right\|^2_{L_2(P^X)} .$$

So we have

$$\begin{aligned}
&\sup_{s\in S_L} P_n\left(Ks_M - Ks\right)\\
&= \sup_{s\in S_L}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right) - (P_n-P)\left(s-s_M\right)^2 - P\left(Ks-Ks_M\right)\right\}\\
&= \sup_{s\in S_L}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right) - (P_n-P)\left(s-s_M\right)^2 - L\right\}\ .
\end{aligned}$$

This is where the hypothesis of consistency in sup-norm plays a role by allowing to neglect the second order term $\sup_{s\in S_L}\left\{(P_n-P)\left(s-s_M\right)^2\right\}$. Indeed, if we assume that with high probability

$$\|s_n - s_M\|_\infty \leq R_{n,D} \ll 1\ ,$$

we can localize our analysis in the ball $B_{(M,L_\infty)}\left(s_M, R_{n,D}\right)$ with high probability instead of the entire M. So S_L should be replaced by $D_L = S_L \cap B_{(M,L_\infty)}\left(s_M, R_{n,D}\right)$ and in Lemma 7.4 we show that the mean of the second order term is negligible, the deviations being handled in the proof by Bousquet's inequality. So let us write

$$\begin{aligned}
\sup_{s\in D_L} P_n\left(Ks_M - Ks\right) &\approx \sup_{s\in D_L}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right) - L\right\}\\
&= \sup_{s\in D_L}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)\right\} - L\ .
\end{aligned}$$

At that stage, a fundamental remark is that $s\in D_L$ implies $2s_M - s \in D_L$, and hence

$$\sup_{s\in D_L}\left\{(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)\right\} = \sup_{s\in D_L}\left|(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)\right|\ . \tag{7.42}$$

Indeed, as we want to bound as sharply as possible the value of C_*, we have to bound from below the right-hand side of the latter inequality, and this is possible thanks to the "linearity" recovered on the functions $\psi_{1,M}\cdot(s_M-s)$, $s\in D_L$. This is indeed one of the main reasons why we expanded the contrast. As a matter of fact, separating the linear term from the second order term allows us to put the absolute values, whereas we unfortunately have

$$\sup_{s\in D_L}\left\{(P_n-P)\left(Ks_M-Ks\right)\right\} \neq \sup_{s\in D_L}\left|(P_n-P)\left(Ks_M-Ks\right)\right|$$

in general. Moreover, linearity is lost on the contrasted functions, even if the model M is linear or affine. So in other situations than regression or M-estimation with regular contrast, if one wants to apply the approach developed in Section 7.3, the first question to address is to understand how to put absolute values inside the supremum of the empirical process, as in (7.42) above, and how to bound from below the considered moment of order one for the supremum of the empirical process. The strict equality in (7.42) could be relaxed, and a manner to relate the one-sided supremum of the empirical process to the two-sided supremum is for instance to control sufficiently sharply the variance of the one-sided supremum of the empirical process. Furthermore, a lower bound for the moment of order one for the supremum of the empirical process can be found in Giné and Koltchinskii [39], Theorem 3.4, under regularly varying entropy bounds, providing a first general answer to this probabilistic question. However, this is done under the assumption of a lower bound on the variance of the elements of the index set that could be restrictive, considering the problem of lower bounds for the empirical excess risk in the general context of M-estimation.

As $D_L \subset S_L$, we get by Cauchy-Schwarz inequality that

$$\sup_{s\in D_L}\left|(P_n-P)\left(\psi_{1,M}\cdot(s_M-s)\right)\right| \leq \sqrt{\sum_{k=1}^{D}\left((P_n-P)\left(\psi_{1,M}\cdot\varphi_k\right)\right)^2} \tag{7.43}$$

where $(\varphi_k)_{k=1}^D$ is an orthonormal basis of M and so, with high probability,

$$\begin{aligned}
&\sup_{s\in D_L} |(P_n - P)(\psi_{1,M}\cdot(s_M - s))| \\
&\approx \mathbb{E}\left[\sup_{s\in D_L} |(P_n - P)(\psi_{1,M}\cdot(s_M - s))|\right] \\
&\leq \mathbb{E}^{1/2}\left[\left(\sup_{s\in D_L} |(P_n - P)(\psi_{1,M}\cdot(s_M - s))|\right)^2\right] \qquad (7.44)\\
&= \sqrt{\frac{LD}{n}}\mathcal{K}_{1,M}\ .
\end{aligned}$$

As we need to prove that

$$\sup_{s\in D_L} |(P_n - P)(\psi_{1,M}\cdot(s_M - s))| \approx \sqrt{\frac{LD}{n}}\mathcal{K}_{1,M}\ ,$$

we should reverse the inequalities (7.43) and (7.44). The first inequality can be reversed if we show that the function in S_L achieving the Cauchy-Schwarz inequality belongs with high probability to $B_{(M,L_\infty)}(s_M, R_{n,D})$. This is done in Lemma 7.3, that strongly relies on the structure of localized basis of the model. To reverse the second inequality, we use the moment inequality established in Corollary 7.2, and the proper lower bound is obtained in Lemma 7.2. Hence, we have with high probability that

$$\sup_{s\in D_L} P_n(Ks_M - Ks) \approx \sqrt{\frac{LD}{n}}\mathcal{K}_{1,M} - L\ .$$

Finally, taking the argument of the maximum over values of L in the right-hand side of the latter inequality we get

$$C_* \approx \frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2\ .$$

In order to fix ideas we obtain in the simple case of homoscedastic regression, taking for instance a model M of histograms defined on a partition $\mathcal{P}$, by using (3.33) given in Section 3.3.3 of Chapter 3 and denoting σ the constant noise level, a rate of order

$$C_* \approx \frac{1}{4}\frac{D}{n}\mathcal{K}_{1,M}^2 = \frac{D\sigma^2}{n} + \frac{1}{n}\sum_{I\in\mathcal{P}} \mathbb{V}\left[\mathbb{E}\left[Y\,|X\right]|X\in I\right]\ .$$

7.4 Probabilistic Tools

7.4.1 Classical tools

We recall here the main probabilistic results that are instrumental in our proofs.
Let us begin with the L_p-version of Hoffmann-Jørgensen's inequality, that can be found for example in [53], Proposition 6.10, p.157.

Theorem 7.3 *For any independent mean zero random variables Y_j, $j = 1, ..., n$ taking values in a Banach space $(\mathcal{B}, \|.\|)$ and satisfying $\mathbb{E}[\|Y_j\|^p] < +\infty$ for some $p \geq 1$, we have*

$$\mathbb{E}^{1/p}\left\|\sum_{j=1}^n Y_j\right\|^p \leq B_p\left(\mathbb{E}\left\|\sum_{j=1}^n Y_j\right\| + \mathbb{E}^{1/p}\left(\max_{1\leq j\leq n}\|Y_j\|\right)^p\right)$$

where B_p is a universal constant depending only on p.

We will use this theorem for $p = 2$ in order to control suprema of empirical processes. In order to be more specific, let $\mathcal{F}$ be a class of measurable functions from a measurable space $\mathcal{Z}$ to $\mathbb{R}$ and $(X_1, ..., X_n)$ be independent variables of common law P taking values in $\mathcal{Z}$. We then denote by $\mathcal{B} = l^\infty(\mathcal{F})$ the space of uniformly bounded functions on $\mathcal{F}$ and, for any $b \in \mathcal{B}$, we set $\|b\| = \sup_{f\in\mathcal{F}} |b(f)|$. Thus $(\mathcal{B}, \|.\|)$ is a Banach space. Indeed we shall apply Theorem 7.3 to the independent random variables, with mean zero and taking values in $\mathcal{B}$, defined by

$$Y_j = \{f(X_j) - Pf,\ f \in \mathcal{F}\} .$$

More precisely, we will use the following result, which is a straightforward application of Theorem 7.3. Denote by

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

the empirical measure associated to the sample $(X_1, ..., X_n)$ and by

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |(P_n - P)(f)|$$

the supremum of the empirical process over $\mathcal{F}$.

Corollary 7.1 *If $\mathcal{F}$ is a class of measurable functions from a measurable space $\mathcal{Z}$ to $\mathbb{R}$ satisfying*

$$\sup_{z\in\mathcal{Z}} \sup_{f\in\mathcal{F}} |f(z) - Pf| = \sup_{f\in\mathcal{F}} \|f - Pf\|_\infty < +\infty$$

and $(X_1, ..., X_n)$ are n i.i.d. random variables taking values in $\mathcal{Z}$, then an absolute constant B_2 exists such that,

$$\mathbb{E}^{1/2}\left[\|P_n - P\|_{\mathcal{F}}^2\right] \leq B_2\left(\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \frac{\sup_{f\in\mathcal{F}} \|f - Pf\|_\infty}{n}\right) . \tag{7.45}$$

Another tool we need is a comparison theorem for Rademacher processes, see Theorem 4.12 of [53]. A function $\varphi : \mathbb{R} \to \mathbb{R}$ is called a contraction if $|\varphi(u) - \varphi(v)| \leq |u - v|$ for all $u, v \in \mathbb{R}$. Moreover, for a subset $T \subset \mathbb{R}^n$ we set

$$\|h(t)\|_T = \|h\|_T = \sup_{t\in T} |h(t)| .$$

Theorem 7.4 *Let $(\varepsilon_1, ..., \varepsilon_n)$ be n i.i.d. Rademacher variables and $F : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ be a convex and increasing function. Furthermore, let $\varphi_i : \mathbb{R} \longrightarrow \mathbb{R}$, $i \leq n$, be contractions such that $\varphi_i(0) = 0$. Then, for any bounded subset $T \subset \mathbb{R}^n$,*

$$\mathbb{E}F\left(\left\|\sum_i \varepsilon_i \varphi_i(t_i)\right\|_T\right) \leq 2\mathbb{E}F\left(\left\|\sum_i \varepsilon_i t_i\right\|_T\right) .$$

The next tool is the well known Bernstein's inequality, that can be found for example in [61], Proposition 2.9.

Theorem 7.5 *(Bernstein's inequality) Let $(X_1, ..., X_n)$ be independent real valued random variables and define*

$$S = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) .$$

Assuming that

$$v = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] < \infty$$

and

$$X_i \leq b \quad a.s.$$

we have, for every $x > 0$,

$$\mathbb{P}\left[|S| \geq \sqrt{2v\frac{x}{n}} + \frac{bx}{3n}\right] \leq 2\exp(-x). \tag{7.46}$$

We now turn to concentration inequalities for the empirical process around its mean. Bousquet's inequality [28] provides optimal constants for the deviations above the mean. Klein-Rio's inequality [41] gives sharp constants for the deviations below the mean, that slightly improves Klein's inequality [42].

Theorem 7.6 *Let* $(\xi_1, ..., \xi_n)$ *be* n *i.i.d. random variables having common law* P *and taking values in a measurable space* $\mathcal{Z}$. *If* $\mathcal{F}$ *is a class of measurable functions from* $\mathcal{Z}$ *to* $\mathbb{R}$ *satisfying*

$$|f(\xi_i) - Pf| \leq b \quad a.s., \text{ for all } f \in \mathcal{F},\ i \leq n,$$

then, by setting

$$\sigma_{\mathcal{F}}^2 = \sup_{f\in\mathcal{F}} \left\{P\left(f^2\right) - (Pf)^2\right\},$$

we have, for all $x \geq 0$,

Bousquet's inequality *:*

$$\mathbb{P}\left[\|P_n - P\|_{\mathcal{F}} - \mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \geq \sqrt{2\left(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right]\right)\frac{x}{n}} + \frac{bx}{3n}\right] \leq \exp(-x) \tag{7.47}$$

and we can deduce that, for all $\varepsilon, x > 0$, *it holds*

$$\mathbb{P}\left[\|P_n - P\|_{\mathcal{F}} - \mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \geq \sqrt{2\sigma_{\mathcal{F}}^2\frac{x}{n}} + \varepsilon\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + \frac{1}{3}\right)\frac{bx}{n}\right] \leq \exp(-x). \tag{7.48}$$

Klein-Rio's inequality :

$$\mathbb{P}\left[\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2\left(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right]\right)\frac{x}{n}} + \frac{bx}{n}\right] \leq \exp(-x) \tag{7.49}$$

and again, we can deduce that, for all $\varepsilon, x > 0$, *it holds*

$$\mathbb{P}\left[\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2\sigma_{\mathcal{F}}^2\frac{x}{n}} + \varepsilon\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + 1\right)\frac{bx}{n}\right] \leq \exp(-x). \tag{7.50}$$

7.4.2 A moment inequality for the supremum of the empirical process

For short let us denote by $Z = \|P_n - P\|_{\mathcal{F}}$ the supremum of the empirical process over a class of functions $\mathcal{F}$. Our goal here is to show that the moments of order 1 and 2 of Z are equivalent, in good cases among which those relevant for our needs. In such cases we would like a result of the form

$$\left|\mathbb{E}[Z] - \mathbb{E}^{1/2}\left[Z^2\right]\right| \ll \mathbb{E}[Z] \tag{7.51}$$

and we can notice that it suffices to show that $\sqrt{\mathbb{V}(Z)} = \sqrt{\mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]} \ll \mathbb{E}[Z]$. In order to derive Inequality (7.51) - see Corollary 7.2 for a precise result - we thus begin by establishing in Theorem 7.7 below a general upper bound on the variance $\mathbb{V}(Z)$ of the supremum of the empirical process over a class of functions $\mathcal{F}$.

Theorem 7.7 *Let $\mathcal{F}$ be a class of uniformly bounded functions from $\mathcal{Z}$ to $\mathbb{R}$. For n i.i.d. random variables $(\xi_1, ..., \xi_n)$ taking values in $\mathcal{Z}$, denote by $Z = \|P_n - P\|_{\mathcal{F}}$ the supremum over the class $\mathcal{F}$ of the empirical process associated to $(\xi_1, ..., \xi_n)$. If σ and b are two positive constants such that $\sigma^2 \geq \sup_{f\in\mathcal{F}} \mathrm{Var}(f)$, $b \geq \sup_{f\in\mathcal{F}} \|f - Pf\|_\infty$, then by setting*

$$A_{n,\sigma,\delta,b} = 2n\sigma^2 \left(1 + \delta^{-1}\right)^{-2} b^{-2}$$

for any $\delta > 0$, we get

$$\mathbb{V}(Z) \leq (\delta \mathbb{E}[Z])^2 + \frac{16\sigma^2}{n} + 4\sqrt{2\pi}\delta \mathbb{E}[Z] \frac{\sigma}{\sqrt{n}} \tag{7.52}$$

$$+ 16 (A_{n,\sigma,\delta,b} + 1) \exp(-A_{n,\sigma,\delta,b}) \frac{\left(1 + \delta^{-1}\right)^2 b^2}{n^2} \tag{7.53}$$

$$+ 8\delta \mathbb{E}[Z] \frac{\left(1 + \delta^{-1}\right) b}{n} \exp(-A_{n,\sigma,\delta,b}) \ . \tag{7.54}$$

In Theorem 7.7, we control the variance of the supremum of the empirical process for a general class of functions $\mathcal{F}$. Indeed, it can be applied to deduce sharp relations between moments of order 1 and 2 of the supremum of the empirical process in various contexts. We use it in Corollary 7.2 under the minimal assumptions pertaining to the situations described in Chapters 3, 5 and 6 so that the control is sharp enough for our needs.

Corollary 7.2 *Under notations of Theorem 7.7, if some $\varkappa_n \in (0, 1)$ exists such that*

$$\varkappa_n^2 \mathbb{E}\left[Z^2\right] \geq \frac{\sigma^2}{n}$$

and

$$\varkappa_n^2 \sqrt{\mathbb{E}[Z^2]} \geq \frac{b}{n}$$

then we have for a numerical constant $A_{1,-}$ ($129 + 4\sqrt{2\pi}$ holds),

$$(1 - \varkappa_n A_{1,-}) \sqrt{\mathbb{E}[Z^2]} \leq \mathbb{E}[Z] \ .$$

We can notice that Hoffman-Jørgensen's inequality - see Theorem 7.3 - can be applied to compare moments of Z, but only up to a universal constant. From this point this point of view, Corollary 7.2 thus provides an improvement over Inequality (7.45) under more restrictive assumptions.
We now establish the proofs of the results stated in Theorem 7.7 and Corollary 7.2.

Proof of Corollary 7.2. We use Theorem 7.7, noticing the fact that

$$\sqrt{\mathbb{E}[Z^2]} - \mathbb{E}[Z] \leq \sqrt{\mathbb{V}(Z)} \ . \tag{7.55}$$

Hence, we shall control the terms given by the right-hand side of the last inequality of Theorem 7.7. We take $\delta = \varkappa_n$, and so

$$(\delta \mathbb{E}[Z])^2 = \varkappa_n^2 (\mathbb{E}[Z])^2 \leq \varkappa_n^2 \mathbb{E}\left[Z^2\right] \ . \tag{7.56}$$

Moreover, we have

$$\begin{aligned} \frac{16\sigma^2}{n} + 4\sqrt{2\pi}\delta \mathbb{E}[Z] \frac{\sigma}{\sqrt{n}} &\leq 16\varkappa_n^2 \mathbb{E}\left[Z^2\right] + 4\sqrt{2\pi}\varkappa_n \sqrt{\mathbb{E}[Z^2]} \times \varkappa_n \sqrt{\mathbb{E}[Z^2]} \\ &\leq \left(16 + 4\sqrt{2\pi}\right) \varkappa_n^2 \mathbb{E}\left[Z^2\right] \ . \end{aligned} \tag{7.57}$$

Hence, using (7.56) and (7.57) we can bound the term given by (7.52), and more precisely,

$$(\delta\mathbb{E}[Z])^2 + \frac{16\sigma^2}{n} + 4\sqrt{2\pi}\delta\mathbb{E}[Z]\frac{\sigma}{\sqrt{n}} \leq \left(17 + 4\sqrt{2\pi}\right)\varepsilon_n^2\mathbb{E}\left[Z^2\right] . \tag{7.58}$$

We turn now to the control of the term given by (7.53). As $A_{n,\sigma,\delta,b} = 2n\sigma^2\left(1+\delta^{-1}\right)^{-2}b^{-2} > 0$ and $\delta = \varkappa_n < 1$, we have

$$\exp\left(-A_{n,\sigma,\delta,b}\right) \leq 1 \quad \text{and} \quad 1+\delta^{-1} < 2\delta^{-1}$$

and so it holds

$$\begin{aligned}
&16\left(A_{n,\sigma,\delta,b}+1\right)\frac{\left(1+\delta^{-1}\right)^2 b^2}{n^2}\exp\left(-A_{n,\sigma,\delta,b}\right)\\
&\leq 16A_{n,\sigma,\delta,b}\frac{\left(1+\delta^{-1}\right)^2 b^2}{n^2} + 16\frac{\left(1+\delta^{-1}\right)^2 b^2}{n^2}\\
&\leq 32\frac{\sigma^2}{n} + 64\frac{\delta^{-2}b^2}{n^2}\\
&\leq 32\varkappa_n^2\mathbb{E}\left[Z^2\right] + 64\varkappa_n^2\mathbb{E}\left[Z^2\right] = 96\varkappa_n^2\mathbb{E}\left[Z^2\right] .
\end{aligned} \tag{7.59}$$

It remains to control the term given by (7.54). We have

$$\begin{aligned}
&8\delta\mathbb{E}[Z]\frac{\left(1+\delta^{-1}\right)b}{n}\exp\left(-A_{n,\sigma,\delta,b}\right)\\
&\leq 16\frac{b}{n}\sqrt{\mathbb{E}\left[Z^2\right]}\\
&\leq 16\varkappa_n^2\mathbb{E}\left[Z^2\right] .
\end{aligned} \tag{7.60}$$

Now, using (7.58), (7.59) and (7.60), we get by Theorem 7.7 that

$$\mathbb{V}(Z) \leq \left(129 + 4\sqrt{2\pi}\right)\varkappa_n^2\mathbb{E}\left[Z^2\right] .$$

for some positive constant A_5 independent of D and n. Combining the last inequality with (7.55) gives the result. ■

We conclude the section by proving our general bound.

Proof of Theorem 7.7. From (7.48) and (7.50) with $\varepsilon = \delta$, we deduce that for any $\delta, x > 0$, it holds

$$\mathbb{P}\left[|Z - \mathbb{E}[Z]| \geq \sqrt{2}\sigma\sqrt{\frac{x}{n}} + \delta\mathbb{E}[Z] + \left(1+\delta^{-1}\right)\frac{bx}{n}\right] \leq 2\exp(-x) . \tag{7.61}$$

Moreover,

$$\mathbb{V}(Z) = \mathbb{E}\left[|Z - \mathbb{E}[Z]|^2\right] = \int_0^{+\infty}\mathbb{P}\left[|Z-\mathbb{E}[Z]| \geq y\right]2ydy \tag{7.62}$$

and we have

$$\int_0^{\delta\mathbb{E}[Z]}\underbrace{\mathbb{P}\left[|Z-\mathbb{E}[Z]| \geq y\right]}_{\leq 1}2ydy \leq (\delta\mathbb{E}[Z])^2 . \tag{7.63}$$

Now, if $x \leq A_{n,\sigma,\delta,b}$ then

$$\begin{aligned}
\sqrt{2}\sigma\sqrt{\frac{x}{n}} + \left(1+\delta^{-1}\right)\frac{bx}{n} &\leq \left(\sqrt{2}\sigma + b\left(1+\delta^{-1}\right)\sqrt{\frac{A_{n,\sigma,\delta,b}}{n}}\right)\sqrt{\frac{x}{n}}\\
&= 2\sqrt{2}\sigma\sqrt{\frac{x}{n}},
\end{aligned}$$

so, by (7.61), we get for $x \leq A_{n,\sigma,\delta,b}$,

$$\mathbb{P}\left[|Z - \mathbb{E}[Z]| \geq 2\sqrt{2}\sigma\sqrt{\frac{x}{n}} + \delta\mathbb{E}[Z]\right] \leq 2\exp(-x)$$

and thus

$$\begin{aligned}
&\int_{\delta\mathbb{E}[Z]}^{2\sqrt{2}\sigma\sqrt{\frac{A_{n,\sigma,\delta,b}}{n}}+\delta\mathbb{E}[Z]} \mathbb{P}[|Z - \mathbb{E}[Z]| \geq y]\, 2y dy \\
&= \int_0^{A_{n,\sigma,\delta,b}} \mathbb{P}\left[|Z - \mathbb{E}[Z]| \geq 2\sqrt{2}\sigma\sqrt{\frac{x}{n}} + \delta\mathbb{E}[Z]\right]\left(2\sqrt{2}\sigma\sqrt{\frac{x}{n}} + \delta\mathbb{E}[Z]\right)\frac{2\sqrt{2}\sigma}{\sqrt{nx}}dx \\
&\leq \int_0^{A_{n,\sigma,\delta,b}} 2\exp(-x)\frac{8\sigma^2}{n}dx + \delta\mathbb{E}[Z]\frac{2\sqrt{2}\sigma}{\sqrt{n}}\int_0^{A_{n,\sigma,\delta,b}} 2\exp(-x)\frac{1}{\sqrt{x}}dx \\
&\leq \frac{16\sigma^2}{n} + 4\sqrt{2\pi}\delta\mathbb{E}[Z]\frac{\sigma}{\sqrt{n}}. \qquad (7.64)
\end{aligned}$$

Next we remark that for $x \geq A_{n,\sigma,\delta,b}$, we have

$$\sqrt{\frac{x}{n}} \leq \sqrt{\frac{x}{A_{n,\sigma,\delta,b}}}\sqrt{\frac{x}{n}} \leq \sqrt{\frac{n}{A_{n,\sigma,\delta,b}}}\frac{x}{n},$$

which yields

$$\begin{aligned}
\sqrt{2}\sigma\sqrt{\frac{x}{n}} + \left(1+\delta^{-1}\right)\frac{bx}{n} &\leq \left(\sqrt{2}\sigma\sqrt{\frac{n}{A_{n,\sigma,\delta,b}}} + \left(1+\delta^{-1}\right)b\right)\frac{x}{n} \\
&= 2\left(1+\delta^{-1}\right)b\frac{x}{n}.
\end{aligned}$$

Hence, we obtain by (7.61) that for $x \geq A_{n,\sigma,\delta,b}$,

$$\mathbb{P}\left[|Z - \mathbb{E}[Z]| \geq 2\left(1+\delta^{-1}\right)b\frac{x}{n} + \delta\mathbb{E}[Z]\right] \leq 2\exp(-x).$$

We also remark that

$$2\sqrt{2}\sigma\sqrt{\frac{A_{n,\sigma,\delta,b}}{n}} + \delta\mathbb{E}[Z] = 2\left(1+\delta^{-1}\right)b\frac{A_{n,\sigma,\delta,b}}{n} + \delta\mathbb{E}[Z].$$

Therefore,

$$\begin{aligned}
&\int_{2\sqrt{2}\sigma\sqrt{\frac{A_{n,\sigma,\delta,b}}{n}}+\delta\mathbb{E}[Z]}^{+\infty} \mathbb{P}[|Z - \mathbb{E}[Z]| \geq y]\, 2y dy \\
&= \int_{2(1+\delta^{-1})b\frac{A_{n,\sigma,\delta,b}}{n}+\delta\mathbb{E}[Z]}^{+\infty} \mathbb{P}[|Z - \mathbb{E}[Z]| \geq y]\, 2y dy \\
&= \int_{A_{n,\sigma,\delta,b}}^{+\infty} \mathbb{P}\left[|Z - \mathbb{E}[Z]| \geq \frac{2\left(1+\delta^{-1}\right)bx}{n} + \delta\mathbb{E}[Z]\right] \times \\
&\qquad \times 2\left(2\left(1+\delta^{-1}\right)b\frac{x}{n} + \delta\mathbb{E}[Z]\right)\frac{2\left(1+\delta^{-1}\right)b}{n}dx \\
&\leq 16\frac{\left(1+\delta^{-1}\right)^2 b^2}{n^2}\int_{A_{n,\sigma,\delta,b}}^{+\infty} x\exp(-x)\,dx + 8\delta\mathbb{E}[Z]\frac{\left(1+\delta^{-1}\right)b}{n}\int_{A_{n,\sigma,\delta,b}}^{+\infty}\exp(-x)\,dx \\
&\leq 16\left(A_{n,\sigma,\delta,b}+1\right)\exp\left(-A_{n,\sigma,\delta,b}\right)\frac{\left(1+\delta^{-1}\right)^2 b^2}{n^2} + 8\delta\mathbb{E}[Z]\frac{\left(1+\delta^{-1}\right)b}{n}\exp\left(-A_{n,\sigma,\delta,b}\right). \qquad (7.65)
\end{aligned}$$

As we divided the integral given in (7.62) in three parts, it suffices to sum bounds given in (7.63), (7.64) and (7.65) to conclude the proof.

■

7.5 Proofs

7.5.1 Proofs of Section 7.2.3

In order to express the quantities of interest in the proofs of Theorems 7.1 and 7.2, we need preliminary definitions. Let $\alpha > 0$ be fixed and for $R_{n,D,\alpha}$ defined in (**A3**), see Section 7.2.1, we set

$$\tilde{R}_{n,D,\alpha} = \max\left\{ R_{n,D,\alpha} \ , \ A_\infty \sqrt{\frac{D \ln n}{n}} \right\} \tag{7.66}$$

where A_∞ is a positive constant to be chosen later. Moreover, we set

$$\nu_n = \max\left\{ \sqrt{\frac{\ln n}{D}}, \ \sqrt{\frac{D \ln n}{n}}, \ R_{n,D,\alpha} \right\} \ . \tag{7.67}$$

Following then heuristics given in Section 7.3.2, our analysis is localized in the subset

$$B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right) := \left\{ s \in M, \|s - s_M\|_\infty \leq \tilde{R}_{n,D,\alpha} \right\}$$

of M.
Let us define several slices of excess risk on the model M : for any $C \geq 0$,

$$\mathcal{F}_C = \left\{ s \in M, \|s - s_M\|_{H,M}^2 \leq C \right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right) \tag{7.68}$$

$$\mathcal{F}_{>C} = \left\{ s \in M, \|s - s_M\|_{H,M}^2 > C \right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right) \tag{7.69}$$

and for any interval $J \subset \mathbb{R}$,

$$\mathcal{F}_J = \left\{ s \in M, \|s - s_M\|_{H,M}^2 \in J \right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right) \ .$$

We also define, for all $L \geq 0$,

$$D_L = \left\{ s \in M, \|s - s_M\|_{H,M}^2 = L \right\} \bigcap B_{(M,L_\infty)}\left(s_M, \tilde{R}_{n,D,\alpha}\right) \ .$$

The set of constants A_2, L_2, A_H and L_H defined in Section 7.1.1 will be denoted by (**RC**). Recall that by assumption (**A4**) given in Section 7.2.2, it holds

$$\|\psi_{1,M}\|_\infty \leq 4A \tag{7.70}$$

and

$$0 < A_3 = \|\psi_{3,M}\|_\infty < +\infty \ . \tag{7.71}$$

By (7.20), the normalized complexity $\mathcal{K}_{1,M}$ defined in Section 7.2.2 satisfies

$$\mathcal{K}_{1,M} = \sqrt{\frac{1}{D} \sum_{k=1}^{D} \mathrm{Var}\left(\psi_{1,M} \cdot \varphi_k\right)}$$

for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $\left(M_0, \|\cdot\|_{H,M}\right)$. Moreover, inequality (7.23) holds under (**A4**) and we have

$$\mathcal{K}_{1,M} \leq A_1 A_H \ . \tag{7.72}$$

By assumption (**A5**) given in Section 7.2.2, we further have

$$\mathcal{K}_{1,M} \geq A_{\mathcal{K}} > 0 \ . \tag{7.73}$$

Furthermore, notice that since $D \leq A_+ n (\ln n)^{-1/2}$, it holds by assumption (**A3**) stated in Section 7.2.1, for all $n \geq n_0 (A_{cons}, L_H)$,

$$0 \leq R_{n,D,\alpha} < L_H^{-1} .$$

Hence, using inequality (8.7), we have for all $n \geq n_0 (A_{cons}, L_H)$ and for any $s \in M$ such that $\|s - s_M\|_\infty \leq R_{n,D,\alpha} \leq L_H^{-1}$,

$$0 < (1 - L_H R_{n,D,\alpha}) \|s - s_M\|_{H,M}^2 \leq P (Ks - Ks_M) \leq (1 + L_H R_{n,D,\alpha}) \|s - s_M\|_{H,M}^2 . \quad (7.74)$$

Finally, when (**A1**) holds - which is the case when (**A2**) holds, see Remark 7.1 -, we have by (7.11),

$$\sup_{t \in M_0,\, \|t\|_{H,M} \leq 1} \|s - s_M\|_\infty \leq A_\Psi \sqrt{D} \quad (7.75)$$

and so, for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $\left(M_0, \|\cdot\|_{H,M}\right)$, it holds for all $k \in \{1, ..., D\}$, since $\|\varphi_k\|_{H,M} = 1$,

$$\|\varphi_k\|_\infty \leq A_\Psi \sqrt{D} . \quad (7.76)$$

Proofs of the theorems

The proof of Theorem **7.1** relies on Lemmas 7.6, 7.7 and 7.8 stated in Section 7.5.2, and that give sharp estimates of suprema of the empirical process on the constrasted functions over slices of interest. We skip its proof as it follows, using Lemmas 7.6, 7.7 and 7.8 stated in Section 7.5.2, from straightforward adaptation of the proof of Theorem 5.1 of Chapter 5 given in Section 5.5.3 of Chapter 5. The proof of Theorem 7.2 follows from straightforward adaptations of the proof of Theorem 3.2 given in Section 3.6.3 of Chapter 4. To fix ideas let us give in the case of Inequality (7.28), the arguments that lead to the use of Lemmas 7.6, 7.7 and 7.8.

Sketch of proof for Inequality (7.28). Let $\alpha > 0$ be fixed and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $(M, \|\cdot\|_2)$ satisfying (**A2**). Since $D \leq A_+ n (\ln n)^{-1/2}$, it holds by assumption (**A3**), for all $n \geq n_0 (A_{cons}, L_H)$,

$$0 \leq R_{n,D,\alpha} < L_H^{-1} .$$

Let $r \in (1, 2]$ to be chosen later and $C, \widetilde{C} > 0$ such that

$$rC = \frac{D}{4n} \mathcal{K}_{1,M}^2 \quad (7.77)$$

and, for all $n \geq n_0 (A_{cons}, L_H)$,

$$\widetilde{C} = (1 - L_H R_{n,D,\alpha}) C > 0 . \quad (7.78)$$

By inequality (7.74), if

$$P (Ks_n - Ks_M) \leq \widetilde{C} \quad \text{and} \quad \|s_n - s_M\|_\infty \leq R_{n,D,\alpha}$$

then

$$\|s_n - s_M\|_{H,M}^2 \leq C ,$$

for all $n \geq n_0(A_{cons}, L_H)$. Hence, by (**A3**) there exists a positive integer n_1 such that it holds, for all $n \geq n_0(A_{cons}, L_H, n_1)$,

$$\begin{aligned}\mathbb{P}\left(P(Ks_n - Ks_M) \leq \widetilde{C}\right) &\leq \mathbb{P}\left(\left\{P(Ks_n - Ks_M) \leq \widetilde{C}\right\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} \\ &\leq \mathbb{P}\left(\left\{\|s_n - s_M\|_{H,M}^2 \leq C\right\} \bigcap \Omega_{\infty,\alpha}\right) + n^{-\alpha} \ . \end{aligned} \tag{7.79}$$

Now, by definition of the slices $\mathcal{F}_C$ and $\mathcal{F}_{>C}$ respectively given in (7.68) and (7.69), and since by definition of $\tilde{R}_{n,D,\alpha}$ given in (7.66) it holds

$$R_{n,D,\alpha} \leq \tilde{R}_{n,D,\alpha} \ ,$$

we can write

$$\begin{aligned}&\mathbb{P}\left(\left\{\|s_n - s_M\|_{H,M}^2 \leq C\right\} \bigcap \Omega_{\infty,\alpha}\right) \\ &\leq \mathbb{P}\left(\inf_{s\in\mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s\in\mathcal{F}_{>C}} P_n(Ks - Ks_M)\right) \\ &\leq \mathbb{P}\left(\inf_{s\in\mathcal{F}_C} P_n(Ks - Ks_M) \leq \inf_{s\in\mathcal{F}_{(C,rC]}} P_n(Ks - Ks_M)\right) \\ &= \mathbb{P}\left(\sup_{s\in\mathcal{F}_C} P_n(Ks_M - Ks) \geq \sup_{s\in\mathcal{F}_{(C,rC]}} P_n(Ks_M - Ks)\right) \ . \end{aligned} \tag{7.80}$$

Now, the proof of Inequality follows from straightforward adaptations of the proof of Inequality (5.32) of Theorem 3.1 given in Section 5.5.3 of Chapter 5.

7.5.2 Technical lemmas

We state here some lemmas needed in the proofs given in Section 7.5.1. First, in Lemmas 7.1, 7.2 and 7.3, we derive some controls, from above and from below, of the empirical process indexed by the "linear parts" of the contrasted functions over slices of interest. Secondly, we give upper bounds in Lemmas 7.4 and 7.5 for the empirical process indexed by the "quadratic parts" of the contrasted functions over slices of interest. And finally, we use all these results in Lemmas 7.6, 7.7 and 7.8 to derive upper and lower bounds for the empirical process indexed by the contrasted functions over slices of interest.

Lemma 7.1 *Assume that (**A1**), (**A4**) and (**A5**) hold. Then for any $\beta > 0$, by setting*

$$\tau_n = L_{A_1, A_{\mathcal{K}}, A_\Psi, \beta}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right) \ ,$$

we have, for any orthonormal basis $(\varphi_k)_{k=1}^D$ of $\left(M_0, \|\cdot\|_{H,M}\right)$,

$$\mathbb{P}\left[\sqrt{\sum_{k=1}^{D} (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} \geq (1 + \tau_n)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} \ .$$

Proof. By Cauchy-Schwarz inequality we have

$$\chi_M := \sqrt{\sum_{k=1}^{D} (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} = \sup_{t\in M_0,\ \|t\|_{H,M}\leq 1} \{|(P_n - P)(\psi_{1,M} \cdot t)|\} \ .$$

Hence, we get by Bousquet's inequality (7.48) applied with $\mathcal{F} = \Big\{\psi_{1,M} \cdot t \ ; \ t \in M_0, \ \ \|t\|_{H,M} \leq 1\Big\}$, for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi_M \geq \sqrt{2\sigma^2 \frac{x}{n}} + (1+\delta)\,\mathbb{E}\left[\chi_M\right] + \left(\frac{1}{3} + \frac{1}{\delta}\right)\frac{bx}{n}\right] \leq \exp\left(-x\right) \tag{7.81}$$

where

$$\sigma^2 \leq \sup_{t \in M_0,\ \|t\|_{H,M} \leq 1} P\left[\left(\psi_{1,M} \cdot s\right)^2\right] \leq \|\psi_{1,M}\|_\infty^2 \sup_{t \in M_0,\ \|t\|_{H,M} \leq 1} \|t\|_2^2 \leq \left(A_1 A_H\right)^2 \qquad \text{by (7.70)}$$

and

$$b \leq \sup_{t \in M_0,\ \|t\|_{H,M} \leq 1} \left\|\psi_{1,M} \cdot t - P\left(\psi_{1,M} \cdot t\right)\right\|_\infty \leq 2A_1 A_\Psi \sqrt{D} \qquad \text{by (7.70) and (7.75).}$$

Moreover, since by identity (7.20) of Section 7.2.2,

$$\mathcal{K}_{1,M} = \sqrt{\frac{1}{D}\sum_{k=1}^{D} \mathrm{Var}\left(\psi_{1,M} \cdot \varphi_k\right)}\ ,$$

we easily see that

$$\mathbb{E}\left[\chi_M\right] \leq \sqrt{\mathbb{E}\left[\chi_M^2\right]} = \sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\ .$$

So, from (7.81) it follows that, for all $x > 0$, $\delta > 0$,

$$\mathbb{P}\left[\chi_M \geq \sqrt{2\left(A_1 A_H\right)^2 \frac{x}{n}} + (1+\delta)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M} + \left(\frac{1}{3} + \frac{1}{\delta}\right)\frac{2A_1 A_\Psi \sqrt{D} x}{n}\right] \leq \exp\left(-x\right)\ . \tag{7.82}$$

Hence, taking $x = \beta \ln n$, $\delta = \frac{\sqrt{\ln n}}{n^{1/4}}$ in (7.82), we can derive by (7.73) that

$$\mathbb{P}\left[\chi_M \geq \left(1 + L_{A_1, A_\mathcal{K}, A_\Psi, \beta}\left(\sqrt{\frac{\ln n}{D}} \vee \frac{\sqrt{\ln n}}{n^{1/4}}\right)\right)\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta}\ ,$$

which gives the result.■

In the next lemma, we state sharp lower bounds for the mean of the supremum of the empirical process on the linear parts of constrasted functions of M belonging to a slice of excess risk. This is done for a model of reasonable dimension.

Lemma 7.2 *Let $r > 1$ and $C > 0$. Assume that (**A2**), (**A4**) and (**A5**) hold and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $\left(M_0, \|\cdot\|_{H,M}\right)$ satisfying (**A2**). If positive constants A_-, A_+, A_l, A_u exist such that*

$$A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- \left(\ln n\right)^2 \quad \text{and} \quad A_l \frac{D}{n} \leq rC \leq A_u \frac{D}{n}\ ,$$

and if the constant A_∞ defined in (7.66) satisfies

$$A_\infty \geq 32 B_2 A_1 A_H \sqrt{2A_u} A_\mathcal{K}^{-1} r_M\left(\varphi\right)\ , \tag{7.83}$$

then, $n \geq n_0\left(A_+, A_-, A_l, A_u, A_1, A_H, A_\mathcal{K}, r_M\left(\varphi\right), A_{cons}, B_2\right)$,

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} \left(P_n - P\right)\left(\psi_{1,M} \cdot \left(s_M - s\right)\right)\right] \geq \left(1 - \frac{L_{A_l, A_u, A_\mathcal{K}, A_1, A_H}}{\sqrt{D}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\ . \tag{7.84}$$

Our argument leading to Lemma 7.2 shows that we have to assume that the constant A_∞ introduced in (7.66) is large enough. In order to prove Lemma 7.2 the following result is needed.

Lemma 7.3 *Let $r > 1$, $\beta > 0$ and $C \geq 0$. Assume that (**A2**), (**A4**) and (**A5**) hold and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $\left(M_0, \|\cdot\|_{H,M}\right)$ satisfying (**A2**). If positive constants A_+, A_- and A_u exist such that*

$$A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2 , \quad rC \leq A_u \frac{D}{n} ,$$

and if

$$A_\infty \geq 16 B_2 A_1 A_H \sqrt{2A_u \beta} A_{\mathcal{K}}^{-1} r_M(\varphi)$$

then, for all $n \geq n_0(A_+, A_-, A_1, A_H, A_{\mathcal{K}}, r_M(\varphi), B_2, \beta)$, it holds

$$\mathbb{P}\left[\max_{k \in \{1,\dots,D\}} \left| \frac{\sqrt{rC}\,(P_n - P)(\psi_{1,M} \cdot \varphi_k)}{\sqrt{\sum_{j=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_j)}} \right| \geq \frac{\tilde{R}_{n,D,\alpha}}{r_M(\varphi)\sqrt{D}} \right] \leq \frac{2D+1}{n^\beta} .$$

Proof of Lemma 7.3. By Cauchy-Schwarz inequality, we get

$$\chi_M = \sqrt{\sum_{k=1}^D (P_n - P)^2 (\psi_{1,M} \cdot \varphi_k)} = \sup_{t \in S_{M_0}} |(P_n - P)(\psi_{1,M} \cdot t)| ,$$

where S_{M_0} is the unit sphere of M_0, that is

$$S_{M_0} = \left\{ t \in M_0 \ ; \ t = \sum_{k=1}^D \beta_k \varphi_k \text{ and } \sqrt{\sum_{k=1}^D \beta_k^2} = 1 \right\} .$$

Thus we can apply Klein-Rio's concentration inequality (7.50) to χ_M by taking $\mathcal{F} = S_{M_0}$ and use the fact that

$$\sup_{t \in S_{M_0}} \|\psi_{1,M} \cdot t - P(\psi_{1,M} \cdot t)\|_\infty \leq 2A_1\sqrt{D} r_M(\varphi) \quad \text{by (7.70) and (\textbf{A2}).} \tag{7.85}$$

$$\sup_{t \in S_{M_0}} \mathrm{Var}(\psi_{1,M} \cdot t) \leq \sup_{t \in S_{M_0}} P(\psi_{1,M} \cdot t)^2 \leq (A_1 A_H)^2 \sup_{t \in S_{M_0}} \|t\|_{H,M}^2 = (A_1 A_H)^2 \quad \text{by (7.70)}$$

and also, by using (7.85) in Inequality (7.45) applied to χ_M, we get that

$$\begin{aligned} \mathbb{E}[\chi_M] &\geq B_2^{-1} \sqrt{\mathbb{E}\left[\chi_M^2\right]} - \frac{2A_1\sqrt{D} r_M(\varphi)}{n} \\ &= B_2^{-1} \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - \frac{2A_1\sqrt{D} r_M(\varphi)}{n} . \end{aligned}$$

We thus obtain by (7.50), for all $\varepsilon, x > 0$,

$$\mathbb{P}\left(\chi_M \leq (1-\varepsilon) B_2^{-1} \sqrt{\frac{D}{n}} \mathcal{K}_{1,M} - \sqrt{2(A_1 A_H)^2 \frac{x}{n}} - \left(1 - \varepsilon + \left(1 + \frac{1}{\varepsilon}\right) x\right) \frac{2A_1\sqrt{D} r_M(\varphi)}{n} \right) \leq \exp(-x) . \tag{7.86}$$

So, by taking $\varepsilon = \frac{1}{2}$ and $x = \beta \ln n$ in (7.86), and by observing that $D \geq A_- (\ln n)^2$, $\mathcal{K}_{1,M} \geq A_{\mathcal{K}} > 0$ by (**A5**), we conclude that, for all $n \geq n_0 (A_-, A_1, A_H, A_{\mathcal{K}}, r_M(\varphi), B_2, \beta)$,

$$\mathbb{P}\left[\chi_M \leq \frac{B_2^{-1}}{8}\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}\right] \leq n^{-\beta} \ . \tag{7.87}$$

Furthermore, combining Bernstein's inequality (7.46), with the observation that we have, for every $k \in \{1, ..., D\}$,

$$\begin{aligned} \|\psi_{1,M} \cdot \varphi_k\|_\infty &\leq A_1\sqrt{D} r_M(\varphi) && \text{by (7.70) and (\mathbf{A2})} \\ P(\psi_{1,M} \cdot \varphi_k)^2 &\leq A_H^2 \|\psi_{1,M}\|_\infty^2 \|\varphi_k\|_{H,M}^2 \leq (A_1 A_H)^2 && \text{by (7.70)} \end{aligned}$$

we get that, for every $x > 0$ and every $k \in \{1, ..., D\}$,

$$\mathbb{P}\left[|(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \geq \sqrt{2(A_1 A_H)^2 \frac{x}{n}} + \frac{A_1\sqrt{D} r_M(\varphi)}{3}\frac{x}{n}\right] \leq 2\exp(-x)$$

and so

$$\mathbb{P}\left[\max_{k\in\{1,...,D\}} |(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \geq \sqrt{2(A_1 A_H)^2 \frac{x}{n}} + \frac{A_1\sqrt{D} r_M(\varphi)}{3}\frac{x}{n}\right] \leq 2D\exp(-x) \ . \tag{7.88}$$

Hence, taking $x = \beta \ln n$ in (7.88), it comes

$$\mathbb{P}\left[\max_{k\in\{1,...,D\}} |(P_n - P)(\psi_{1,M} \cdot \varphi_k)| \geq \sqrt{\frac{2(A_1 A_H)^2 \beta \ln n}{n}} + \frac{A_1\sqrt{D} r_M(\varphi)\beta \ln n}{3n}\right] \leq \frac{2D}{n^\beta} \ , \tag{7.89}$$

then, by using (7.87) and (7.89), we get for all $n \geq n_0 (A_-, A_1, A_H, A_{\mathcal{K}}, r_M(\varphi), B_2, \beta)$,

$$\mathbb{P}\left[\max_{k\in\{1,...,D\}} \left|\frac{\sqrt{rC}(P_n - P)(\psi_{1,M} \cdot \varphi_k)}{\chi_M}\right| \geq \frac{8B_2\sqrt{rC}}{\sqrt{\frac{D}{n}}\mathcal{K}_{1,M}}\left(\sqrt{\frac{2(A_1 A_H)^2 \beta \ln n}{n}} + \frac{A_1\sqrt{D} r_M(\varphi)\beta \ln n}{3n}\right)\right]$$
$$\leq \frac{2D+1}{n^\beta} \ .$$

Finally, as $A_+ \frac{n}{(\ln n)^2} \geq D$ we have, for all $n \geq n_0 (A_+, A_1, A_H, r_M(\varphi), \beta)$,

$$\frac{A_1\sqrt{D} r_M(\varphi)\beta \ln n}{3n} \leq \sqrt{\frac{2(A_1 A_H)^2 \beta \ln n}{n}}$$

and we can check that, since $rC \leq A_u \frac{D}{n}$ and $\mathcal{K}_{1,M} \geq A_{\mathcal{K}} > 0$, if

$$A_\infty \geq 16 B_2 A_1 A_H \sqrt{2A_u \beta} A_{\mathcal{K}}^{-1} r_M(\varphi)$$

then, for all $n \geq n_0 (A_+, A_-, A_1, A_H, A_{\mathcal{K}}, r_M(\varphi), B_2, \beta)$,

$$\mathbb{P}\left[\max_{k\in\{1,...,D\}} \left|\frac{\sqrt{rC}(P_n - P)(\psi_{1,M} \cdot \varphi_k)}{\chi_M}\right| \geq \frac{A_\infty}{r_M(\varphi)}\sqrt{\frac{\ln n}{n}}\right] \leq \frac{2D+1}{n^\beta} \ ,$$

which readily gives the result. ■

We are now ready to prove the lower bound (7.84) for the expected value of the largest increment of the empirical process over $\mathcal{F}_{(C,rC]}$.

Proof of Lemma 7.2. Let us begin with the lower bound of $\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2$, a result that will be need further in the proof. Introduce

$$s_{CS}=s_M+\sum_{k=1}^{D}\beta_{k,n}\varphi_k\in M\ ,$$

where for all $k\in\{1,...,D\}$,

$$\beta_{k,n}=\frac{\sqrt{rC}\left(P_n-P\right)\left(\psi_{1,M}\cdot\varphi_k\right)}{\sqrt{\sum_{j=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_j\right)}}\ ,$$

Observe that

$$\left\|s_{CS}\right\|_{H,M}^2=rC\ . \tag{7.90}$$

We also set

$$\tilde{\Omega}=\left\{\max_{k\in\{1,...,D\}}\left|\beta_{k,n}\right|\leq\frac{\tilde{R}_{n,D,\alpha}}{r_M\left(\varphi\right)\sqrt{D}}\right\}\ .$$

By Lemma 7.3 we have for all $\beta>0$, if

$$A_\infty\geq 16B_2A_1A_H\sqrt{2A_u\beta}A_{\mathcal{K}}^{-1}r_M\left(\varphi\right)\ ,$$

for all $n\geq n_0\left(A_+,A_-,A_1,A_H,A_{\mathcal{K}},r_M\left(\varphi\right),B_2,\beta\right)$,

$$\mathbb{P}\left(\tilde{\Omega}\right)\geq 1-\frac{2D+1}{n^\beta}\ . \tag{7.91}$$

Moreover, by (**A2**), we get on the event $\tilde{\Omega}$,

$$\left\|\sum_{k=1}^{D}\beta_{k,n}\varphi_k\right\|_\infty\leq\tilde{R}_{n,D,\alpha}\ ,$$

and so, on $\tilde{\Omega}$,

$$s_{CS}\in\mathcal{F}_{(C,rC]}\ . \tag{7.92}$$

As a consequence, by (7.92) it holds

$$\begin{aligned}
&\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2\\
&\geq\mathbb{E}^{\frac{1}{2}}\left[\left(\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(\sum_{k=1}^{D}\beta_{k,n}\varphi_k\right)\right)\right)^2 1_{\tilde{\Omega}}\right]\\
&=\sqrt{rC}\sqrt{\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)1_{\tilde{\Omega}}\right]}\ .
\end{aligned}\tag{7.93}$$

Furthermore, since by (**A2**) $\left\|\varphi_k\right\|_\infty\leq\sqrt{D}r_M\left(\varphi\right)$ for all $k\in\{1,...,D\}$, we have

$$\begin{aligned}
\left|\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right|&\leq D\max_{k=1,...,D}\left|\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right|\\
&\leq 4D\max_{k=1,...,D}\left\|\psi_{1,M}\cdot\varphi_k\right\|_\infty^2\\
&\leq A_1^2D^2r_M^2\left(\varphi\right)
\end{aligned}$$

and it ensures

$$\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)1_{\tilde{\Omega}}\right]\geq\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)\right]-A_1^2D^2r_M^2\left(\varphi\right)\mathbb{P}\left[\left(\tilde{\Omega}\right)^c\right]\ . \tag{7.94}$$

Comparing inequality (7.94) with (3.176) and using (7.91), we obtain the following lower bound, for all $n\geq n_0\left(A_+,A_-,A_1,A_H,A_{\mathcal{K}},r_M\left(\varphi\right),B_2,\beta\right)$,

$$\begin{aligned}\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2&\geq\sqrt{rC}\sqrt{\mathbb{E}\left[\left(\sum_{k=1}^{D}\left(P_n-P\right)^2\left(\psi_{1,M}\cdot\varphi_k\right)\right)\right]}\\&\quad-A_1r_M\left(\varphi\right)D\sqrt{rC}\sqrt{\mathbb{P}\left[\left(\tilde{\Omega}\right)^c\right]}\\&\geq\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}-A_1r_M\left(\varphi\right)D\sqrt{rC}\sqrt{\frac{2D+1}{n^\beta}}\ .\end{aligned} \tag{7.95}$$

We take $\beta=4$, and so we must have

$$A_\infty\geq 32B_2A_1A_H\sqrt{2A_u}A_{\mathcal{K}}^{-1}r_M\left(\varphi\right)\ .$$

Since $D\leq A_+n\left(\ln n\right)^{-2}$ and $\mathcal{K}_{1,M}\geq A_{\mathcal{K}}>0$, we get, for all $n\geq n_0\left(A_1,A_+,A_{\mathcal{K}},r_M\left(\varphi\right)\right)$,

$$A_1r_M\left(\varphi\right)D\sqrt{rC}\sqrt{\frac{2D+1}{n^\beta}}\leq\frac{1}{\sqrt{D}}\times\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M} \tag{7.96}$$

and so, by using (7.95) and (7.96), for all $n\geq n_0\left(A_+,A_-,A_1,A_H,A_{\mathcal{K}},r_M\left(\varphi\right),B_2\right)$,

$$\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2\geq\left(1-\frac{1}{\sqrt{D}}\right)\sqrt{\frac{rCD}{n}}\mathcal{K}_{1,M}\ . \tag{7.97}$$

Now, as $D\geq A_-\left(\ln n\right)^2$ we have for all $n\geq n_0\left(A_-\right)$, $D^{-1/2}\leq 1/2$. Moreover, we have $\mathcal{K}_{1,M}\geq A_{\mathcal{K}}>0$ and $rC\geq A_lDn^{-1}$, so we finally deduce from (7.97) that, for all $n\geq n_0\left(A_+,A_-,A_l,A_1,A_H,A_{\mathcal{K}},r_M\left(\varphi\right),B_2\right)$,

$$\mathbb{E}^{\frac{1}{2}}\left(\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right)^2\geq\frac{A_{\mathcal{K}}}{2}\sqrt{A_l}\frac{D}{n}\ . \tag{7.98}$$

We turn now to the lower bound of $\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right]$. First observe that $s\in\mathcal{F}_{(C,rC]}$ implies that

$$2s_M-s=s_M-\left(s-s_M\right)\in\mathcal{F}_{(C,rC]}\ ,$$

so that

$$\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}}\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right]=\mathbb{E}\left[\sup_{s\in\mathcal{F}_{(C,rC]}}\left|\left(P_n-P\right)\left(\psi_{1,M}\cdot\left(s_M-s\right)\right)\right|\right]\ . \tag{7.99}$$

In the next step, we apply Corollary 7.2. More precisely, using notations of Corollary 7.2, we set

$$\mathcal{F}=\left\{\psi_{1,M}\cdot\left(s_M-s\right),s\in\mathcal{F}_{(C,rC]}\right\}$$

and

$$Z = \sup_{s \in \mathcal{F}_{(C,rC]}} \left| (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right) \right| \ .$$

Now, since for all $n \geq n_0\left(A_+, A_-, A_\infty, A_{cons}\right)$ we have $\tilde{R}_{n,D,\alpha} \leq 1$, we get by (7.70), for all $n \geq n_0\left(A_+, A_-, A_\infty, A_{cons}\right)$,

$$\sup_{f \in \mathcal{F}} \|f - Pf\|_\infty = 2 \sup_{s \in \mathcal{F}_{(C,rC]}} \|\psi_{1,M} \cdot (s_M - s)\|_\infty \leq 2A_1 \tilde{R}_{n,D,\alpha} \leq 2A_1$$

we set $b = 2A_1$. Since we assume that $rC \leq A_u \frac{D}{n}$, it moreover holds by (7.70),

$$\sup_{f \in \mathcal{F}} \mathrm{Var}\left(f\right) \leq \sup_{s \in \mathcal{F}_{(C,rC]}} P\left(\psi_{1,M} \cdot (s_M - s)\right)^2 \leq \left(A_1 A_H\right)^2 rC \leq \left(A_1 A_H\right)^2 A_u \frac{D}{n}$$

and so we set $\sigma^2 = \left(A_1 A_H\right)^2 A_u \frac{D}{n}$. Now, by (7.98) we have, for all $n \geq n_0\left(A_+, A_-, A_l, A_1, A_H, A_\mathcal{K}, r_M\left(\varphi\right), B_2\right)$,

$$\sqrt{\mathbb{E}\left[Z^2\right]} \geq \frac{A_\mathcal{K}}{2} \sqrt{A_l} \frac{D}{n} \ . \tag{7.100}$$

Hence, a positive constant $L_{A_l,A_u,A_\mathcal{K},A_1,A_H}$ ($\max\left(2A_1 A_H A_\mathcal{K}^{-1} \sqrt{A_u/A_l} \ ; \ 2\sqrt{A_1} A_\mathcal{K}^{-1} A_l^{-1/4}\right)$ holds) exists such that, by setting

$$\varkappa_n = \frac{L_{A_l,A_u,A_\mathcal{K},A_1,A_H}}{\sqrt{D}}$$

we can get, using (7.100), that,
for all $n \geq n_0\left(A_+, A_-, A_l, A_u, A_1, A_H, A_\mathcal{K}, r_M\left(\varphi\right), A_{cons}, B_2\right)$,

$$\varkappa_n^2 \mathbb{E}\left[Z^2\right] \geq \frac{\sigma^2}{n} \ ,$$

$$\varkappa_n^2 \sqrt{\mathbb{E}\left[Z^2\right]} \geq \frac{b}{n} \ .$$

Furthermore, since $D \geq A_- \left(\ln n\right)^2$, we have for all $n \geq n_0\left(A_-, A_l, A_u, A_\mathcal{K}, A_1, A_H\right)$,

$$\varkappa_n \in (0,1) \ .$$

So, using (7.99) and Corollary 7.2, it holds for all $n \geq n_0\left(A_+, A_-, A_l, A_u, A_1, A_H, A_\mathcal{K}, r_M\left(\varphi\right), A_{cons}, B_2\right)$,

$$\begin{aligned} &\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right)\right] \\ &\geq \left(1 - \frac{L_{A_l,A_u,A_\mathcal{K},A_1,A_H}}{\sqrt{D}}\right) \mathbb{E}^{\frac{1}{2}}\left(\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right)\right)^2 \ . \end{aligned} \tag{7.101}$$

Finally, by comparing (7.97) and (7.101), we can deduce that
for all $n \geq n_0\left(A_+, A_-, A_l, A_u, A_1, A_H, A_\mathcal{K}, r_M\left(\varphi\right), A_{cons}, B_2\right)$,

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_{(C,rC]}} (P_n - P)\left(\psi_{1,M} \cdot (s_M - s)\right)\right] \geq \left(1 - \frac{L_{A_l,A_u,A_\mathcal{K},A_1,A_H}}{\sqrt{D}}\right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M}$$

and so (7.84) is proved. ■
Let us now turn to the control of second order terms appearing in the expansion of the regular contrast K, see (7.4).

Lemma 7.4 *Let $C \geq 0$ and $A_+ > 0$. Assume that*

$$A_+ \frac{n}{(\ln n)^2} \geq D \ .$$

It holds, for all $n \geq n_0 (A_+, A_{cons}, A_\infty, A_3, L_2)$,

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_C} |(P_n - P)(\psi_2 \circ (\psi_{3,M} \cdot (s - s_M)))|\right] \leq 8 L_2 A_3^2 A_H \tilde{R}_{n,D,\alpha} \sqrt{\frac{CD}{n}} \ .$$

Proof. We define the Rademacher process $\mathcal{R}_n$ on a class $\mathcal{F}$ of measurable functions from $\mathcal{Z}$ to $\mathbb{R}$, to be

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(\xi_i) \ , \quad f \in \mathcal{F}$$

where ε_i are independent Rademacher random variables also independent from the ξ_i. By the usual symmetrization argument we have

$$\mathbb{E}\left[\sup_{s \in \mathcal{F}_C} \left|(P_n - P)\left(\psi_2 \circ \left(\frac{s - s_M}{s_M}\right)\right)\right|\right] \leq 2\mathbb{E}\left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (\psi_{3,M} \cdot (s - s_M)))|\right] \ . \quad (7.102)$$

Recall that $A_3 := \|\psi_{3,M}\|_\infty > 0$. As $A_+ \frac{n}{(\ln n)^2} \geq D$, we have for all $n \geq n_0 (A_+, A_{cons}, A_\infty, A_3, L_2)$,

$$\tilde{R}_{n,D,\alpha} \leq (A_3 L_2)^{-1} \ .$$

Hence, for all $n \geq n_0 (A_+, A_{cons}, A_\infty, A_3, L_2)$ and for all $(x, y) \in \left[-A_3 \tilde{R}_{n,D,\alpha},\ A_3 \tilde{R}_{n,D,\alpha}\right]^2$, it holds from (7.5), since $A_3 \tilde{R}_{n,D,\alpha}$

$$|\psi_2(x) - \psi_2(y)| \leq L_2 A_3 \tilde{R}_{n,D,\alpha} |x - y| \ . \quad (7.103)$$

We define now the following real-valued function ρ,

$$\rho(x) = \begin{cases} \left(L_2 A_3 \tilde{R}_{n,D,\alpha}\right)^{-1} \psi_2(x) & \text{if } x \in \left[-A_3 \tilde{R}_{n,D,\alpha},\ A_3 \tilde{R}_{n,D,\alpha}\right] \\ \left(L_2 A_3 \tilde{R}_{n,D,\alpha}\right)^{-1} \psi_2\left(-A_3 \tilde{R}_{n,D,\alpha}\right) & \text{if } x \leq -A_{\min}^{-1} \tilde{R}_{n,D,\alpha} \\ \left(L_2 A_3 \tilde{R}_{n,D,\alpha}\right)^{-1} \psi_2\left(A_3 \tilde{R}_{n,D,\alpha}\right) & \text{if } x \geq A_{\min}^{-1} \tilde{R}_{n,D,\alpha} \end{cases}$$

and since $\rho(0) = \psi_2(0) = 0$, it follows from (7.103) that ρ is a contraction mapping for all $n \geq n_0 (A_+, A_{cons}, A_\infty, A_3, L_2)$. Then, taking the expectation with respect to the Rademacher variables, we then get for all $n \geq n_0 (A_+, A_{cons}, A_\infty, A_3, L_2)$,

$$\begin{aligned} &\mathbb{E}_\varepsilon\left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(\psi_2 \circ (\psi_{3,M} \cdot (s - s_M)))|\right] \\ &= L_2 A_3 \tilde{R}_{n,D,\alpha} \mathbb{E}_\varepsilon\left[\sup_{s \in \mathcal{F}_C} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i \rho\left((\psi_{3,M} \cdot (s - s_M))(\xi_i)\right)\right|\right] \end{aligned} \quad (7.104)$$

We can now apply Theorem 7.4 to get for all $n \geq n_0 (A_+, A_{cons}, A_\infty, A_3, L_2)$,

$$\begin{aligned} \mathbb{E}_\varepsilon\left[\sup_{s \in \mathcal{F}_C} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i \rho\left((\psi_{3,M} \cdot (s - s_M))(\xi_i)\right)\right|\right] &\leq 2\mathbb{E}_\varepsilon\left[\sup_{s \in \mathcal{F}_C} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i (\psi_{3,M} \cdot (s - s_M))(\xi_i)\right|\right] \\ &= 2\mathbb{E}_\varepsilon\left[\sup_{s \in \mathcal{F}_C} |\mathcal{R}_n(\psi_{3,M} \cdot (s - s_M))|\right] \end{aligned} \quad (7.105)$$

and so we derive successively the following upper bounds in mean, for all $n \geq n_0\left(A_+, A_{cons}, A_\infty, A_3, L_2\right)$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2\circ\left(\psi_{3,M}\cdot(s-s_M)\right)\right)\right|\right] = \mathbb{E}\left[\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_2\circ\left(\psi_{3,M}\cdot(s-s_M)\right)\right)\right|\right]\right] \\
&\leq L_2A_3\tilde{R}_{n,D,\alpha}\mathbb{E}\left[\mathbb{E}_\varepsilon\left[\sup_{s\in\mathcal{F}_C}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\rho\left(\psi_{3,M}\cdot(s-s_M)\left(\xi_i\right)\right)\right|\right]\right] \qquad \text{by (7.104)} \\
&\leq 2L_2A_3\tilde{R}_{n,D,\alpha}\mathbb{E}\left[\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_{3,M}\cdot(s-s_M)\right)\right|\right] \qquad \text{by (7.105)} \\
&\leq 2L_2A_3\tilde{R}_{n,D,\alpha}\sqrt{\mathbb{E}\left[\left(\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_{3,M}\cdot(s-s_M)\right)\right|\right)^2\right]}\,. \qquad (7.106)
\end{aligned}$$

We consider now an orthonormal basis of $\left(M_0, \|\cdot\|_{H,M}\right)$ and denote it by $(\varphi_k)_{k=1}^D$. Whence

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[\left(\sup_{s\in\mathcal{F}_C}\left|\mathcal{R}_n\left(\psi_{3,M}\cdot(s-s_M)\right)\right|\right)^2\right]} \\
&\leq \sqrt{\mathbb{E}\left[\left(\sup\left\{\left|\sum_{k=1}^D\beta_k\mathcal{R}_n\left(\psi_{3,M}\cdot\varphi_k\right)\right|;\ \sum_{k=1}^D\beta_k^2\leq C\right\}\right)^2\right]} \\
&= \sqrt{C}\sqrt{\mathbb{E}\left[\sum_{k=1}^D\left(\mathcal{R}_n\left(\psi_{3,M}\cdot\varphi_k\right)\right)^2\right]} \\
&= \sqrt{\frac{C\sum_{k=1}^D P\left(\psi_{3,M}\cdot\varphi_k\right)^2}{n}} \leq A_3A_H\sqrt{\frac{CD}{n}}\,, \qquad (7.107)
\end{aligned}$$

and the result follows by injecting (7.106) and (7.107) in (7.102). ■

In the following Lemma, we provide uniform upper bounds for the supremum of the empirical process of second order terms in the contrast expansion when the considered slices are not too small. We skip its proof as it follows, using Lemma 7.4, from straightforward adaptations of the proof of Lemma 3.11 given in Section 3.6.4 of Chapter 3.

Lemma 7.5 *Let $A_+, A_-, A_l, \beta, C_- > 0$, and assume (**H3**) and (**H5**). Then if $C_- \geq A_l\frac{D}{n}$ and $A_+n\left(\ln n\right)^{-2} \geq D \geq A_-\left(\ln n\right)^2$, then a positive constant $L_{A_-,A_l,L_2,A_3,A_H,\beta}$ exists such that, for all $n \geq n_0\left(A_+, A_l, A_{cons}, A_\infty, A_3, L_2\right)$,*

$$\mathbb{P}\left[\forall C > C_-,\quad \sup_{s\in\mathcal{F}_C}\left|\left(P_n - P\right)\left(\psi_2\circ(s-s_M)\right)\right| \leq L_{A_-,A_l,L_2,A_3,A_H,\beta}\sqrt{\frac{CD}{n}}\tilde{R}_{n,D,\alpha}\right] \geq 1 - n^{-\beta}\,.$$

Having controlled the residual empirical process driven by the remainder terms in the contrast, and having proved sharp bounds for the expectation of the increments of the main empirical process on the slices, it remains to combine the above lemmas in order to establish the probability estimates controlling the empirical excess risk on the slices. We skip the proofs of the three next lemmas as they follow, using Lemmas 7.1, 7.2 and 7.5, from straightforward adaptations of Lemmas 5.13, 5.14 and 5.15 given in Section 5.5.5 of Chapter 5.

Lemma 7.6 *Let $\beta, A_-, A_+, A_l, C > 0$. Assume that (**A1**), (**A3**), (**A4**) and (**A5**) hold. A positive constant A_4 exists, only depending on $A_1, A_{\mathcal{K}}, A_\Psi, A_-, A_l, L_2, A_3, A_H, \beta$, such that, if*

$$A_l \frac{D}{n} \leq C \leq \frac{1}{4}\left(1 + A_4 \nu_n\right)^2 \frac{D}{n} \mathcal{K}_{1,M}^2 \quad \text{and} \quad A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D}},\ \sqrt{\frac{D \ln n}{n}},\ R_{n,D,\alpha}\right\}$ is defined in (7.67), then
for all $n \geq n_0\left(A_+, A_l, A_\infty, A_{cons}, A_\infty, A_3, L_2, n_1, \alpha\right)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_C} P_n\left(Ks_M - Ks\right) \geq \left(1 + L_{A_1, A_{\mathcal{K}}, A_\Psi, A_-, A_l, L_2, A_3, A_H, \beta} \times \nu_n\right) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C\right] \leq 2n^{-\beta} \ .$$

Lemma 7.7 *Let $\beta, A_-, A_+, C \geq 0$. Assume that (**A1**), (**A3**), (**A4**) and (**A5**) hold. A positive constant A_5, depending on $A_1, A_{\mathcal{K}}, A_\Psi, A_-, A_l, L_2, A_3, A_H, \beta$, exists such that, if it holds*

$$C \geq \frac{1}{4}\left(1 + A_5 \nu_n\right)^2 \frac{D}{n} \mathcal{K}_{1,M}^2 \quad \text{and} \quad A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D}},\ \sqrt{\frac{D \ln n}{n}},\ R_{n,D,\alpha}\right\}$ is defined in (7.67),
then for all $n \geq n_0\left(A_+, A_l, A_\infty, A_{cons}, A_\infty, A_3, L_2, n_1, \alpha\right)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \left(1 + A_5 \nu_n\right) \sqrt{\frac{CD}{n}} \mathcal{K}_{1,M} - C\right] \leq 2n^{-\beta} \ .$$

Moreover, when we only assume $C \geq 0$, we have for all $n \geq n_0\left(A_+, A_\infty, A_{cons}, A_\infty, A_3, L_2, n_1, \alpha\right)$,

$$\mathbb{P}\left[\sup_{s \in \mathcal{F}_{>C}} P_n\left(Ks_M - Ks\right) \geq \frac{1}{4}\left(1 + A_5 \nu_n\right)^2 \frac{D}{n} \mathcal{K}_{1,M}^2\right] \leq 2n^{-\beta} \ . \tag{7.108}$$

Lemma 7.8 *Let $r > 1$ and $C, \beta > 0$. Assume that (**A2**), (**A3**), (**A4**) and (**A5**) hold and let $\varphi = (\varphi_k)_{k=1}^D$ be an orthonormal basis of $\left(M, \|\cdot\|_{H,M}\right)$ satisfying (**A2**). If positive constants A_-, A_+, A_l, A_u exist such that*

$$A_+ \frac{n}{(\ln n)^2} \geq D \geq A_- (\ln n)^2 \quad \text{and} \quad A_l \frac{D}{n} \leq rC \leq A_u \frac{D}{n} \ ,$$

and if the constant A_∞ defined in (7.66) satisfies

$$A_\infty \geq 32 B_2 A_1 A_H \sqrt{2A_u} A_{\mathcal{K}}^{-1} r_M\left(\varphi\right) \ ,$$

then for all $n \geq n_0\left(A_-, A_+, A_u, A_l, A_1, A_H, A_\infty, A_{cons}, B_2, r_M\left(\varphi\right), A_{\mathcal{K}}, n_1, \alpha\right)$,

$$\mathbb{P}\left(\sup_{s \in \mathcal{F}_{(C, rC]}} P_n\left(Ks_M - Ks\right) \leq \left(1 - L_{A_-, A_l, A_u, A, A_\infty, A_{\mathcal{K}}, A_1, A_H, r_M(\varphi), \beta} \times \nu_n\right) \sqrt{\frac{rCD}{n}} \mathcal{K}_{1,M} - rC\right) \leq 2n^{-\beta} \ ,$$

where $\nu_n = \max\left\{\sqrt{\frac{\ln n}{D}},\ \sqrt{\frac{D \ln n}{n}},\ R_{n,D,\alpha}\right\}$ is defined in (7.67).

Chapitre 8

Slope heuristics in regular contrast estimation

This chapter is devoted to the validation of slope heuristics in the general context of regular contrast estimation described in Chapter 2. This heuristics, that claims the existence in a model selection via penalization framework of an optimal penalty and a minimal penalty such that the optimal penalty is twice the minimal one, was first formulated by Birgé and Massart in [23] in a generalized Gaussian linear model setting. The formulation of the slope heuristics has then been extended by Arlot and Massart to a general M-estimation setting and the authors validate it in a heteroscedastic with random design regression framework, considering suitable linear histogram models, see [10]. For a more precise presentation of this subject and definition of the slope heuristics, we refer to the introduction of Chapter 4 and also to Section 4.2.2 of the latter chapter.

As shown in Chapter 2, regular contrast estimation contains at least heteroscedastic regression on finite dimensional linear models, least-squares estimation of density on affine models with finite dimensional underlying vector space and maximum likelihood estimation of density on histograms. The slope heuristics will be derived under the assumption that the considered collection of models has a number of elements which is polynomial in the amount of data. This condition is also assumed in [10] and to our best knowledge, the slope heuristics considering more general collections of models has only been proved in a Gaussian setting by Birgé and Massart [23].

The Chapter is organized as follows. Section 8.1 is devoted to the statistical framework. We state in Section 8.2 our theoretical results and we comment on the set of assumptions needed. The proofs are postponed to the end of the chapter.

8.1 Statistical framework

Let $(\mathcal{Z}, \mathcal{T})$ be a measurable space, P an unknown probability measure on $(\mathcal{Z}, \mathcal{T})$ and $\mathcal{S}$ a set of measurable functions from $(\mathcal{Z}, \mathcal{T})$ to $\mathbb{R}$. We also define $\xi_1, ..., \xi_n$ to be n independent random variables with common law P on $(\mathcal{Z}, \mathcal{T})$ and we take a generic random variable ξ of law P, independent of the sample $(\xi_1, ..., \xi_n)$. We consider a contrast K on $\mathcal{S}$ for the law P, that is a functional from $\mathcal{S}$ to $L_1^-(P)$,

$$K : \left\{ \begin{array}{l} \mathcal{S} \longrightarrow L_1^-(P) \\ s \longmapsto (Ks : z \longmapsto (Ks)(z)) \end{array} \right. ,$$

such that there exists a unique element $s_* \in \mathcal{S}$, called the target, satisfying

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) \quad \text{and} \quad P(Ks_*) < +\infty . \tag{8.1}$$

Let us recall the definition of the space $L_1^-(P)$ of real-valued measurable functions on $(\mathcal{Z},\mathcal{T})$ whose negative part is of finite expectation with respect to P. The positive part of a real number $x \in \mathbb{R}$ is denoted $(x)_+ := \max\{x\ ,\ 0\} \geq 0$ and its negative part is $(x)_- := (-x)_+ = \max\{-x\ ,\ 0\} \geq 0$. We naturally extend these definitions to real-valued functions, and for a function f defined from $\mathcal{Z}$ to $\mathbb{R}$,

$$(f)_+ : z \in \mathcal{Z} \longmapsto (f(z))_+ \quad , \quad (f)_- : z \in \mathcal{Z} \longmapsto (f(z))_- \ .$$

Then, $L_1^-(P)$ is defined to be

$$L_1^-(P) = \left\{ f : \mathcal{Z} \to \mathbb{R} \ \ \mathcal{T}\text{-measurable}\ ;\ P(f)_- < +\infty \right\} \ .$$

Notice that expectation with respect to P is well-defined on $L_1^-(P)$, by writing for any $f \in L_1^-(P)$,

$$Pf := P(f)_+ - P(f)_- \in \overline{\mathbb{R}}\ ,$$

where $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. We also set $L_1(P)$ the set of integrable real-valued functions for the law P,

$$L_1(P) = \{ f : (\mathcal{Z},\mathcal{T}) \to \mathbb{R}\ ;\ P|f| < +\infty \}\ ,$$

$L_2(P)$ is the set of square integrable real-valued functions for the law P,

$$L_2(P) = \left\{ f : (\mathcal{Z},\mathcal{T}) \to \mathbb{R}\ ;\ \|f\|_2 := \sqrt{P(f)^2} < +\infty \right\}\ ,$$

and $L_\infty(P)$ is the set real-valued functions essentially bounded on $\mathcal{Z}$ with respect to the law P,

$$L_\infty(P) := \{ s : (\mathcal{Z},\mathcal{T}) \to \mathbb{R}\ ;\ \|s\|_\infty := \mathrm{essup}_{z\in\mathcal{Z}}(|s(z)|) < +\infty \}\ .$$

The target s_* is, according to (8.1), the minimizer of the risk $P(Ks)$ over the set $\mathcal{S}$. It is an unknown quantity as it depends on the law P. Our goal is to estimate the target s_* by using the sample $(\xi_1,...,\xi_n)$. To that end, we consider a finite collection of models $\mathcal{M}_n$ such that for all $M \in \mathcal{M}_n$, we have $M \subset \mathcal{S} \bigcap L_\infty(P)$. For each $M \in \mathcal{M}_n$ we take a M-estimator s_n on M, assumed to exist but non necessarily unique, satisfying

$$s_n \in \arg\min_{s\in M} P_n(Ks) \quad \text{with} \quad P_n(Ks_n) < +\infty \ a.s., \tag{8.2}$$

where

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$$

is the empirical distribution of the sample $(\xi_1,...,\xi_n)$.
Moreover, we assume that the contrast K is uniformly regular for the models M of the collection $\mathcal{M}_n$ and the law P, which means that the contrast is regular for each $M \in \mathcal{M}_n$ and that the constants involved in the definition of its regularity, which is given in Section 2.2 of Chapter 2, are uniform over the collection $\mathcal{M}_n$. Let us now explicitly state our definition of contrast in our model selection framework. First, for every $M \in \mathcal{M}_n$ there exists a unique projection s_M of s_* on M, defined to be

$$s_M = \arg\min_{s\in M} P(Ks) \quad \text{with} \quad P(Ks_M) < +\infty\ . \tag{8.3}$$

In addition, consider a function ψ_2 defined on a subset $\mathcal{D}_2 \subseteq \mathbb{R}$ such that $0 \in \mathring{\mathcal{D}}_2$, where $\mathring{\mathcal{D}}_2$ denotes the interior of $\mathcal{D}_2$, $\psi_2(\mathcal{D}_2) \subseteq \overline{\mathbb{R}}$ and $\psi_2(0) = 0$. For all $M \in \mathcal{M}_n$, for all $s \in M$ and P-almost all $z \in \mathcal{Z}$, the following expansion hold,

$$Ks(z) - Ks_M(z) = \psi_0^s + \psi_{1,M}(z)(s - s_M)(z) + \psi_2(\psi_{3,M}(z)(s - s_M)(z))\ , \tag{8.4}$$

where ψ_0^s is a constant depending on s but not on z, $\psi_{1,M}$ and $\psi_{3,M}$ are functions defined on $\mathcal{Z}$ not depending on s and not identically equal to 0 satisfying $\psi_{1,M} \in L_2(P)$, $\psi_{3,M} \in L_2(P)$. Moreover, there exists $L_2 > 0$ such that for all $\delta \in \left[0, L_2^{-1}\right]$, it holds $[-\delta, \delta] \subset \mathcal{D}_2$ and for all $(x, y) \in [-\delta, \delta]$,

$$|\psi_2(x) - \psi_2(y)| \leq L_2 \delta |x - y| \ . \tag{8.5}$$

Thirdly, for each $M \in \mathcal{M}_n$, we denote

$$\widetilde{M_0} := \text{Span}\{s - s_M \ ; \ s \in M\}$$

and there exists positive constants $A_H, L_H > 0$ such that for all $M \in \mathcal{M}_n$, there exists an Hilbertian norm $\|\cdot\|_{H,M}$ on $\widetilde{M_0}$ satisfying

$$\|\cdot\|_2 \leq A_H \|\cdot\|_{H,M} \tag{8.6}$$

and for all $\delta \in \left[0, L_H^{-1}\right]$, for all $s \in M$ such that $\|s - s_M\|_\infty \leq \delta \leq L_H^{-1}$, it holds

$$(1 - L_H \delta) \|s - s_M\|_{H,M}^2 \leq P(Ks - Ks_M) \leq (1 + L_H \delta) \|s - s_M\|_{H,M}^2 \ . \tag{8.7}$$

We further assume in this chapter that for every $M \in \mathcal{M}_n$, the model M is affine, which means that

$$M_0 := \{s - s_M \ ; \ s \in M\} \tag{8.8}$$

is a linear vector space, and we demand that M_0 has a finite linear dimension that we denote D_M. This gives $M_0 = \widetilde{M_0}$ and so, $\|\cdot\|_{H,M}$ is an Hilbertian norm on M_0. By abuse, M_0 is also called a model for all $M \in \mathcal{M}_n$.

For comments on the previous definition and its relation to situations treated in Chapters 4, 5 and 6, see Section 8.2.3 below.

Let us now define the model selection procedure. We measure the performance of the M-estimators by their excess risk,

$$l(s_*, s_n(M)) := P(Ks_n(M) - Ks_*) \ . \tag{8.9}$$

Moreover, we have

$$l(s_*, s_n(M)) = l(s_*, s_M) + l(s_M, s_n(M)) \ ,$$

where the quantity

$$l(s_*, s_M) := P(Ks_M - Ks_*)$$

is called the bias of the model M and $l(s_M, s_n(M)) := P(Ks_n(M) - Ks_M) \geq 0$ is the excess risk of the M-estimator $s_n(M)$ on M. Notice that we prove sharp bounds for the latter quantity in Chapter 7.

Given the collection of models $\mathcal{M}_n$, an oracle model M_* is defined to be

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \{l(s_*, s_n(M))\} \tag{8.10}$$

and the associated oracle estimator $s_n(M_*)$ thus achieves the best performance in terms of excess risk among the collection $\{s_n(M) ; M \in \mathcal{M}_n\}$. Unfortunately, the oracle model is unknown as it depends on the unknown law P of the data, and we propose to estimate it by a model selection procedure via penalization. Given some known penalty pen, that is a function from $\mathcal{M}_n$ to $\mathbb{R}_+$, we thus consider the following data-dependent model, also called selected model,

$$\widehat{M} \in \arg\min_{M \in \mathcal{M}_n} \{P_n(Ks_n(M)) + \text{pen}(M)\} \ . \tag{8.11}$$

The primary goal for a statistician using a model selection via penalization procedure is then to find - and use - a good penalty, such that the selected model $\widehat{M}$ satisfies an oracle inequality. By an oracle inequality, we mean here an inequality of the form

$$l\left(s_*, s_n\left(\widehat{M}\right)\right) \leq C \times \ell\left(s_*, s_n\left(M_*\right)\right) \ ,$$

with some positive constant C as close to one as possible and with high probability, typically more than $1 - Ln^{-2}$ for some positive constant L.
Our aim in this chapter is to theoretically validate the slope heuristics as it has been formulated in [10] by Arlot and Massart, and thus to prove the existence, under suitable assumptions that are stated in Section 8.2.1 below, of an optimal penalty and a minimal one such the optimal penalty is close to two times the minimal one. More precisely, our aim is to theoretically justify the following facts:

(i) If a penalty $\text{pen} : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ is such that, for all model $M \in \mathcal{M}_n$,

$$\text{pen}\left(M\right) \leq \left(1 - \delta\right) \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$$

with $\delta > 0$, then the dimension of the selected model $\widehat{M}$ is "very large" and the excess risk of the selected estimator $s_n\left(\widehat{M}\right)$ is "much larger" than the excess risk of the oracle.

(ii) If $\text{pen} \approx \left(1 + \delta\right) \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$ with $\delta > 0$, then the corresponding model selection procedure satisfies an oracle inequality with a leading constant $C\left(\delta\right) < +\infty$ and the dimension of the selected model is "not too large". Moreover,

$$\text{pen}_{\text{opt}} \approx 2\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$$

is an optimal penalty.

The mean of the empirical excess risk on M, when M varies in $\mathcal{M}_n$, is thus conjectured to be the maximal value of penalty under which the model selection procedure totally misbehaves. It is called the *minimal penalty,* denoted by pen_{min} :

$$\text{for all } M \in \mathcal{M}_n, \qquad \text{pen}_{\text{min}}\left(M\right) = \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \geq 0 \ .$$

The optimal penalty is then close to two times the minimal one,

$$\text{pen}_{\text{opt}} \approx 2\,\text{pen}_{\text{min}} \ .$$

8.2 Results

8.2.1 Set of assumptions

We state now the set of assumptions needed to derive the results of Section 8.2.2 below.

Set of assumptions : (SA)

(P1) Polynomial complexity of $\mathcal{M}_n$: $\text{Card}\left(\mathcal{M}_n\right) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(P2) Upper bound on dimensions of models in $\mathcal{M}_n$: there exists a positive constant $A_{\mathcal{M},+}$ such that for every $M \in \mathcal{M}_n$, $1 \leq D_M \leq A_{\mathcal{M},+} n \left(\ln n\right)^{-2} \leq n$.

(P3) Richness of $\mathcal{M}_n$: there exist $M_0, M_1 \in \mathcal{M}_n$ such that $D_{M_0} \in [\sqrt{n}, c_{rich}\sqrt{n}]$ and $D_{M_1} \geq A_{rich} n \left(\ln n\right)^{-2}$.

(Aurc) As described in Section 8.1, where we defined the constants L_2, A_H and L_H, the contrast K is uniformly regular over the collection $\mathcal{M}_n$.

(Ab) Coefficients in the contrast expansions over the models of the collection $\mathcal{M}_n$ are uniformly bounded on $\mathcal{Z}$: There exist two positive constants A_1, A_3 such that for all $M \in \mathcal{M}_n$,

$$\|\psi_{1,M}\|_\infty \leq A_1 \ , \tag{8.12}$$

$$\|\psi_{3,M}\|_\infty \leq A_3 \ . \tag{8.13}$$

(AKl) The normalized complexities are bounded from below, uniformly over the collection $\mathcal{M}_n$: There exists $A_{\mathcal{K}} > 0$ such that for all $M \in \mathcal{M}_n$,

$$\mathcal{K}_{1,M} \geq A_{\mathcal{K}} > 0 \ . \tag{8.14}$$

($\mathbf{Ap}_u$) The bias decreases as a power of D_M: there exist $\beta_+ > 0$ and $C_+ > 0$ such that

$$\ell\left(s_*, s_M\right) \leq C_+ D_M^{-\beta_+} \ .$$

(Alb) Each model is provided with a localized basis: there exists a constant $r_{\mathcal{M}}$ such that for each $M \in \mathcal{M}_n$ one can find an orthonormal basis $(\varphi_k)_{k=1}^{D_M}$ of $\left(M_0, \|\cdot\|_{H,M}\right)$ satisfying that, for all $(\beta_k)_{k=1}^{D_M} \in \mathbb{R}^{D_M}$,

$$\left\| \sum_{k=1}^{D_M} \beta_k \varphi_k \right\|_\infty \leq r_{\mathcal{M}} \sqrt{D_M} \left|\beta\right|_\infty \ ,$$

where $\left|\beta\right|_\infty = \max\left\{|\beta_k| \ ; \ k \in \{1, ..., D_M\}\right\}$.

(Aeu) The $L_2\left(P\right)$-norms of the empirical excess risks are uniformly bounded from above: there exist $A_{eu} > 0$ and $\alpha_{eu} \geq 0$ such that for all $M \in \mathcal{M}_n$,

$$0 \leq \mathbb{E}^{1/2}\left[\left(P_n\left(Ks_M - Ks_n\left(M\right)\right)\right)^2\right] \leq A_{eu} n^{\alpha_{eu}} \ .$$

($\mathbf{Ac}_\infty$) Consistency in sup-norm of the M-estimators: an event Ω_∞ of probability at least $1 - n^{-\alpha_s}$, where $\alpha_s = \max\left\{2\alpha_{eu} + 3 \ ; \ 2 + \alpha_{\mathcal{M}}\right\}$ a positive constant A_{cons}, a positive integer n_1 and a collection of positive numbers $\left(R_{n,D_M}\right)_{M \in \mathcal{M}_n}$ exist, such that

$$\sup_{M \in \mathcal{M}_n} R_{n,D_M} \leq \frac{A_{cons}}{\sqrt{\ln n}} \tag{8.15}$$

and for all $M \in \mathcal{M}_n$ it holds on Ω_∞, for all $n \geq n_1$,

$$\|s_n\left(M\right) - s_M\|_\infty \leq R_{n,D_M} \ . \tag{8.16}$$

(Abv) Uniform margin conditions hold between the target and its projection onto the models of the collection: There exists $A_{bv} > 0$ such that for all $M \in \mathcal{M}_n$,

$$\mathrm{Var}\left(Ks_M - Ks_*\right) \leq A_{bv} \times \ell\left(s_*, s_M\right) \ . \tag{8.17}$$

(Abu) The contrasted projections, centered by the contrasted target are uniformly bounded on $\mathcal{Z}$: There exists $A_{bu} > 0$ such that for all $M \in \mathcal{M}_n$,

$$\|Ks_M - Ks_*\|_\infty \leq A_{bu} \ . \tag{8.18}$$

Some comments on these assumptions can be found in Section 8.2.3.

8.2.2 Theorems

Theorem 8.1 *Under the set of assumptions (**SA**) of Section 8.2.1, for $A_{\text{pen}} \in [0,1)$ and $A_p > 0$, we assume that with probability at least $1 - A_p n^{-2}$ we have*

$$0 \leq \text{pen}\left(M_1\right) \leq A_{\text{pen}} \mathbb{E}\left[P_n\left(K s_M - K s_n\left(M_1\right)\right)\right] \ , \tag{8.19}$$

*where the model M_1 is defined in assumption (**P3**) of (**SA**). Then there exist two positive constants A_1, A_2 independent of n such that, with probability at least $1 - A_1 n^{-2}$, we have, for all $n \geq n_0\left(\left(\boldsymbol{SA}\right), A_{\text{pen}}\right)$,*

$$D_{\widehat{M}} \geq A_2 n \ln\left(n\right)^{-2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \ln\left(n\right) \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{8.20}$$

Thus, Theorem 8.1 justifies the first part (**i**) of the slope heuristics exposed in Section 8.1. As a matter of fact, it shows that there exists a level such that if the penalty is smaller than this level for one of the largest models - namely M_1 in the statement of the theorem -, then the dimension of the output is among the largest dimensions of the collection and the excess risk of the selected estimator is much bigger than the excess risk of the oracle. Moreover, this level is given by the mean of the empirical excess risk of the least-squares estimator on each model. The following theorem validates the second part of the slope heuristics.

Theorem 8.2 *Assume that the general set of assumptions (**SA**) of Section 8.2.1 hold. Moreover, for some $\delta \in [0,1)$ and $A_p, A_r > 0$, assume that an event of probability at least $1 - A_p n^{-2}$ exists on which, for every model $M \in \mathcal{M}_n$ such that $D_M \geq A_{\mathcal{M},+}\left(\ln n\right)^3$, it holds*

$$\left(2-\delta\right) \mathbb{E}\left[P_n\left(K s_M - K s_n\left(M\right)\right)\right] \leq \text{pen}\left(M\right) \leq \left(2+\delta\right) \mathbb{E}\left[P_n\left(K s_M - K s_n\left(M\right)\right)\right] \tag{8.21}$$

together with

$$\text{pen}\left(M\right) \leq A_r \frac{\left(\ln n\right)^3}{n} \tag{8.22}$$

*for every model $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+}\left(\ln n\right)^3$. Then, for $\frac{1}{2} > \eta > \left(1-\beta_+\right)_+ / 2$, there exists a positive constant A_3 only depending on $c_{\mathcal{M}}$ given in (**SA**) and on A_p, a positive constant A_4 only depending on constants in the set of assumptions (**SA**), a positive constant A_5 only depending on constants in the set of assumptions (**SA**) and on A_r and a sequence*

$$\theta_n = A_4 \sup_{M \in \mathcal{M}_n} \left\{\varepsilon_n\left(M\right), \ A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_M \leq n^{\eta+1/2}\right\} \leq \frac{A_4\left(1 \vee \sqrt{A_{cons}}\right)}{\left(\ln n\right)^{1/4}} \tag{8.23}$$

such that with probability at least $1 - A_3 n^{-2}$, it holds for all $n \geq n_0\left(\left(\boldsymbol{SA}\right), \eta, \delta\right)$,

$$D_{\widehat{M}} \leq n^{\eta+1/2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\left(\left(\ln n\right)^{-2} + \theta_n\right)}{\left(1-\delta\right)^2}\right) \ell\left(s_*, s_n\left(M_*\right)\right) + A_5 \frac{\left(\ln n\right)^3}{n} \ . \tag{8.24}$$

Assume that in addition, the following assumption holds,

(Ap) *The bias decreases like a power of D_M : there exist $\beta_- \geq \beta_+ > 0$ and $C_+, C_- > 0$ such that*

$$C_- D_M^{-\beta_-} \leq \ell\left(s_*, s_M\right) \leq C_+ D_M^{-\beta_+} \ .$$

Then it holds with probability at least $1 - A_3 n^{-2}$*, for all* $n \geq n_0 \left((\boldsymbol{SA}), C_-, \beta_-, \eta, \delta \right)$*,*

$$A_{\mathcal{M},+} \left(\ln n \right)^{-3} \leq D_{\widehat{M}} \leq n^{\eta + 1/2}$$

and

$$\ell \left(s_*, s_n \left(\widehat{M} \right) \right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{(1-\delta)^2} \right) \ell \left(s_*, s_n \left(M_* \right) \right) \ . \tag{8.25}$$

The quantity $\varepsilon_n (M)$ used in (8.23) controls the deviations of the true and empirical excess risks on the model M and is more precisely defined in Remark 8.1 above. From Theorems 8.1 and 8.2, we identify the minimal penalty with the mean of the empirical excess risk on each model,

$$\mathrm{pen}_{\min} (M) = \mathbb{E} \left[P_n \left(K s_M - K s_n (M) \right) \right] \ .$$

Moreover, Theorem 8.2 states in particular that if the penalty is close to two times the minimal procedure, then the selected estimator satisfies a pathwise oracle inequality with constant almost one, and so the model selection procedure is approximately optimal. If we just assume that the bias of the models decrease at least polynomially with the dimension of the models as it is required in $(\mathbf{Ap}_u)$, then we need to take into account an additional term in the right-hand side of the oracle inequality (8.24), which is proportional to $(\ln n)^3 / n$ and corresponds in the proof of Theorem 8.2 to a uniform upper bound of the model selection criterion on small models of dimension less than $A_{\mathcal{M},+} (\ln n)^3$. Moreover, in this case the dimension of the selected model is much smaller than the largest models as it is smaller than n^λ, for a suitable $\lambda \in (0,1)$. If in addition, we assume that the bias of the models decrease like a power of their dimension, as stated in assumption $(\mathbf{Ap})$ where a polynomial lower bound is required on the bias as well as the same upper bound as in $(\mathbf{Ap}_u)$, then there is no residual term in the nearly optimal pathwise oracle inequality given in (8.25) and the excess risk of the selected estimator is equivalent to the excess risk of the oracle. Moreover, the dimension of the selected model is then larger than $A_{\mathcal{M},+} (\ln n)^3$.

8.2.3 Comments on the set of assumptions

Comments on (P1), (P2), (P3): assumption $(\mathbf{P1})$ states that the number of elements in the collection $\mathcal{M}_n$ is at most polynomial in the amount of data. This assumption allows us to sum the probabilities of deviation of the quantities of interest over the collection $\mathcal{M}_n$. Under $(\mathbf{P1})$, a good criterion uses a penalty function such that the empirical criterion plus this penalty gives an (asymptotically) unbiased estimation of the risk on each model. For more general situations than $(\mathbf{P1})$, the penalty term should take into account the number of elements in the collection $\mathcal{M}_n$, as explained for instance in [61]. Assumption $(\mathbf{P2})$ imposes an upper bound on the dimensions of the considered models and is not too restrictive, as it allows to deal with models than are of the dimension of the amount of data within a power of a logarithmic factor. In assumption $(\mathbf{P3})$, we ask for the existence of a model in $\mathcal{M}_n$ of dimension of order $\sqrt{n}$, and another model of dimension among the largest possible. This is unavoidable to assume the existence of a large and a reasonably large model to show the jump in the dimension of the selected model in regard of the value of the penalty term.

Comments on $(\mathbf{Ap}_u)$ and $(\mathbf{Ap})$: We assume in $(\mathbf{Ap}_u)$ that the quality of approximation of the models of the collection $\mathcal{M}_n$ is good enough in terms of risk. More precisely, we require that the bias of the models are smaller than a power of their dimensions. We recall in Chapter 4 that this is the case considering suitable histogram models if the target is α-Hölderian, and this is again the case when if the models are piecewise polynomials models of uniformly bounded degrees defined on suitable partitions, and if the target belongs to a Besov space of suitable regularity in function of the maximal possible degree of the piecewise polynomials. Assumption $(\mathbf{Ap})$ asks moreover that the bias of the models are bounded from below by a power of their

dimension, which is a quite classical assumption when one wants to deal with the optimality of a model selection procedure, and it has ever been expressed by Stone [69], Burman [29], Arlot [5], [7] and Arlot and Massart in [10]. The latter authors prove in [10] that this is satisfied by considering suitable histogram models when the target is a non-constant α-Hölderian function. We show that this is again the case in maximum likelihood estimation of density, see Section 5.3.3 of Chapter 5, for suitable histogram models and again when the density to be estimated is a non-constant α-Hölderian function.

Comments on (Aurc), (Ab), (AKl), (Alb), ($\mathbf{Ac}_\infty$): As more precisely stated in Remark 8.1 below, these assumptions allow to apply Theorem 7.1 of Chapter 7 with $\alpha = \alpha_s$, where α_s is given in ($\mathbf{Ac}_\infty$). Assumptions **(Aurc), (Ab)** and **(AKl)** are satisfied in heteroscedastic regression, when the data is uniformly bounded from above, the noise level is uniformly bounded away from zero, and the projections of the target onto the models are uniformly bounded from above in the sup-norm, see Chapter 4. However, considering histograms or more general piecewise polynomials defined on suitable lower-regular partitions, assumptions **(Aurc), (Ab), (AKl), (Alb)** and ($\mathbf{Ac}_\infty$) are satisfied if the data is uniformly bounded from above, and if the noise level is uniformly bounded away from zero. Moreover, assumptions **(Aurc), (Ab), (AKl), (Alb)** and ($\mathbf{Ac}_\infty$) are satisfied in maximum likelihood estimation of density on histograms if the considered partitions are lower-regular for the unknown law of data and if the density to be estimated is uniformly bounded from above and uniformly bounded away from zero.

Comments on (Abv), (Abu): these assumptions allow to control the quantity $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$ in Lemma 8.2, by applying Bernstein's inequality. Assumption **(Abv)** states the existence of a uniform margin relation for the projections of the target. This is satisfied in least-squares regression when the projections are assumed to be uniformly bounded in sup-norm, which is the case if the data is uniformly bounded on $\mathcal{Z}$ and if the considered models are histogram or piecewise polynomial models, see Section 2.1.2 of Chapter 2. From the latter section we also easily deduce that **(Abv)** is satisfied in maximum likelihood estimation of density on histograms when the target is uniformly bounded from above and uniformly bounded away from zero. Again, assumption **(Abu)** is satisfied in regression when the data is uniformly bounded on $\mathcal{Z}$ and if the projections of the target are uniformly bounded in sup-norm. In maximum likelihood estimation of density, it is the case when the target is uniformly bounded from above and uniformly bounded away from zero.

Comments on (Aeu): these assumption permits us to derive sharp bounds for the mean of the empirical excess risk on models not too small, see Lemma 8.1 below. It is satisfied in bounded heteroscedastic regression for $\alpha_{eu} = 0$, see Lemma 4.1 of Chapter 4, and in maximum likelihood estimation of density on histograms defined on a lower-regular partition, for any $\alpha_{eu} > 0$, see Lemma 5.6 of Chapter 5. This assumption could be avoided if we achieve concentration bounds for the empirical excess risk of the form: there exist $n_c \in \mathbb{N}$, $x_0 \geq 0$ and $A_0 > 0$ such that for all $n \geq n_c$ and for every $x \geq x_0$, there exists $B_{n,x} > 0$ such that

$$\mathbb{P}\left(P_n\left(Ks_M - Ks_n\right) \geq B_{n,x}\right) \leq A_0 \exp\left(-x\right) \ .$$

Indeed, if $\left(B_{n,x}\right)_{x \geq x_0}$ are sufficiently small, this would lead to (**Aeu**). We believe that this would be possible to adapt the proof of Theorem 7.2 of Chapter 7 in order to obtain such bounds, if we assume that the same type of concentration bounds hold for the quantity $\left\|s_n - s_M\right\|_\infty$. By a careful look of the proof of Theorem 7.2, we believe that the form of $B_{n,x}$ should be

$$B_{n,x} = B_n\left(\sqrt{x} + x\right) \ ,$$

with $B_n > 0$ depending on the constants of the problem and on the number of data. This work is still in progress.

Remark 8.1 *Assume (**P2**), (**Aurc**), (**Ab**), (**AKl**), (**Alb**) and (**Ac**$_\infty$) of the set of assumptions (**SA**) and consider $M \in \mathcal{M}_n$ such that*

$$A_{\mathcal{M},+} (\ln n)^2 \leq D_M \ .$$

Let

$$\alpha_s = \max \{2\alpha_{eu} + 3 \ ; \ 2 + \alpha_{\mathcal{M}}\} > 0 \ . \tag{8.26}$$

*Notice that conditions of Theorem 7.1 are satisfied and that we can apply this theorem with $A_- = A_+ = A_{\mathcal{M},+} > 0$ and $\alpha = \alpha_s > 0$. Hence, a positive finite constant A_0 exists, only depending on $\alpha_s, A_{\mathcal{M},+}$ and on $L_2, A_H, L_H, r_{\mathcal{M}}, A_1$ and $A_{\mathcal{K}}$ defined in the set of assumptions (**SA**), such that by setting*

$$\varepsilon_n (M) = A_0 \max \left\{ \left(\frac{\ln n}{D_M} \right)^{1/4} , \ \left(\frac{D_M \ln n}{n} \right)^{1/4} , \ \sqrt{R_{n,D_M}} \right\} \ , \tag{8.27}$$

we have for all $n \geq n_0 ((\mathbf{SA}))$,

$$\mathbb{P} \left[(1 - \varepsilon_n) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \leq P (Ks_n - Ks_M) \leq (1 - \varepsilon_n) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 10 n^{-\alpha_s} \ , \tag{8.28}$$

$$\mathbb{P} \left[\left(1 - \varepsilon_n^2\right) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \leq P_n (Ks_M - Ks_n) \leq \left(1 + \varepsilon_n^2\right) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5 n^{-\alpha_s} \ , \tag{8.29}$$

*Notice that the constant A_0 in (8.27) is independent of M when M varies in $\mathcal{M}_n$ and satisfies $A_{\mathcal{M},+} (\ln n)^2 \leq D_M$. Morever, if it holds (**P2**), (**Aurc**), (**Ab**), (**Alb**) and (**Ac**$_\infty$) then the conditions of Theorem 7.2 are satisfied for every $M \in \mathcal{M}_n$, since (**Alb**) implies assumption (**A1**) of Chapter 7. Hence, a positive finite constant A_u exists, only depending on $\alpha_s, A_{\mathcal{M},+}$ and on the constants $L_2, A_H, L_H, A_\Psi, A_1$ and $A_{\mathcal{K}}$ defined in (**SA**), such that for all $n \geq n_0 (A_{cons}, n_1)$,*

$$\mathbb{P} \left[P (Ks_n - Ks_M) \geq A_u \frac{D_M \vee \ln n}{n} \right] \leq 3 n^{-\alpha_s} \tag{8.30}$$

and

$$\mathbb{P} \left[P_n (Ks_M - Ks_n) \geq A_u \frac{D_M \vee \ln n}{n} \right] \leq 3 n^{-\alpha_s} \ . \tag{8.31}$$

Again, notice that the constant A_u is independent of M when M varies in $\mathcal{M}_n$.

8.3 Proofs

Before stating the proofs of Theorems 8.2 and 8.1, we need two technical lemmas. In the first lemma, we intend to evaluate the minimal penalty $\mathbb{E} [P_n (Ks_M - Ks_n (M))]$ for models of dimension not too large and not too small.

Lemma 8.1 *Assume (**P2**), (**Aurc**), (**Ab**), (**AKl**), (**Alb**), (**Ac**$_\infty$) and (**Aeu**) of the set of assumptions (**SA**) defined in Section 8.2.1. Then, for every model $M \in \mathcal{M}_n$ of dimension D_M such that*

$$0 < A_{\mathcal{M},+} (\ln n)^2 \leq D_M \ ,$$

we have for all $n \geq n_0 ((\mathbf{SA}))$,

$$\left(1 - 2\varepsilon_n^2 (M)\right) \frac{D_M}{4n} \mathcal{K}_{1,M}^2 \leq \mathbb{E} [P_n (Ks_M - Ks_n (M))] \tag{8.32}$$

$$\leq \left(1 + 2\varepsilon_n^2 (M)\right) \frac{D_M}{4n} \mathcal{K}_{1,M}^2 \ , \tag{8.33}$$

where $\varepsilon_n (M) = A_0 \max \left\{ \left(\frac{\ln n}{D_M} \right)^{1/4} ; \left(\frac{D_M \ln n}{n} \right)^{1/4} ; \sqrt{R_{n,D_M}} \right\}$ is defined in Remark 8.1.

Proof. Let $M \in \mathcal{M}_n$ satisfying $D_M \geq A_{\mathcal{M},+} (\ln n)^2$. As explained in Remark 8.1, under assumptions of Lemma 8.1 we can apply Theorem 7.1 with $A_- = A_+ = A_{\mathcal{M},+}$ and $\alpha = \alpha_s$, where α_s is given by (8.26). For all $n \geq n_0((\mathbf{SA}))$, we thus have on an event $\Omega_1(M)$ of probability at least $1 - 5n^{-\alpha_s}$,

$$(1 - \varepsilon_n(M)) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \leq P_n(Ks_M - Ks_n(M)) \leq (1 + \varepsilon_n(M)) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 , \tag{8.34}$$

where

$$\varepsilon_n(M) = A_0 \max\left\{ \left(\frac{\ln n}{D_M}\right)^{1/4} ; \left(\frac{D_M \ln n}{n}\right)^{1/4} ; \sqrt{R_{n,D_M}} \right\} \geq A_0 n^{-1/8} . \tag{8.35}$$

We also have

$$\begin{aligned} &\mathbb{E}[P_n(Ks_M - Ks_n(M))] \\ &= \mathbb{E}\left[P_n(Ks_M - Ks_n(M)) \mathbf{1}_{\Omega_1(M)}\right] + \mathbb{E}\left[P_n(Ks_M - Ks_n(M)) \mathbf{1}_{(\Omega_1(M))^c}\right] . \end{aligned} \tag{8.36}$$

Hence, as $D_M \geq 1$, it comes from (**AKl**), (**Aeu**) and (8.35) that for all $n \geq n_0(A_0, A_{\mathcal{K}}, A_{eu})$,

$$\begin{aligned} 0 &\leq \mathbb{E}\left[P_n(Ks_M - Ks_n(M)) \mathbf{1}_{(\Omega_1(M))^c}\right] \\ &\leq \mathbb{E}^{1/2}\left[(P_n(Ks_M - Ks_n(M)))^2\right] \sqrt{1 - \mathbb{P}(\Omega_1(M))} \\ &\leq A_{eu} n^{\alpha_{eu}} \sqrt{5n^{-\alpha_s}} \leq \frac{1}{2} \varepsilon_n^2(M) \frac{D_M}{4n} \mathcal{K}_{1,M}^2 . \end{aligned} \tag{8.37}$$

Moreover, we have $\varepsilon_n(M) < 1$ for all $n \geq n_0(A_0, A_{\mathcal{M},+}, A_{cons})$, so by (8.34),

$$0 < \left(1 - 5n^{-\alpha_s}\right)\left(1 - \varepsilon_n^2(M)\right) \frac{D_M}{4n} \mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[P_n(Ks_M - Ks_n(M)) \mathbf{1}_{\Omega_1(M)}\right] \tag{8.38}$$

$$\leq \left(1 + \varepsilon_n^2(M)\right) \frac{D_M}{4n} \mathcal{K}_{1,M}^2 . \tag{8.39}$$

Finally, since we have by (8.35), for all $n \geq n_0(A_0)$,

$$\left(1 - 5n^{-\alpha_s}\right)\left(1 - \varepsilon_n^2(M)\right) > 1 - \frac{3}{2} \varepsilon_n^2(M) ,$$

the result follows by using (8.37), (8.38) and (8.39) in (8.36). ■

Lemma 8.2 *Let $\alpha > 0$. Assume that (**Abv**) and (**Abu**) of Section 8.2.1 are satisfied. Then a positive constant A_d exists, depending only in A, $A_{\mathcal{M},+}$, $\sigma_{\min}$ and α such that, by setting $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$, we have for all $M \in \mathcal{M}_n$,*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq A_d \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n}\right)\right) \leq 2n^{-\alpha} . \tag{8.40}$$

*If moreover, assumptions (**P2**), (**Ab**), (**An**), (**Alb**) and (**Ac**$_\infty$) of the general set of assumptions defined in Section 8.2.1 hold, then for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^2 \leq D_M$ and for all $n \geq n_0((\boldsymbol{SA}))$, we have*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + A_d \frac{\ln n}{\sqrt{D_M}} \mathbb{E}[p_2(M)]\right) \leq 2n^{-\alpha} , \tag{8.41}$$

where $p_2(M) := P_n(Ks_M - Ks_n(M)) \geq 0$.

Proof. We set

$$A_d = \max\left\{\sqrt{2A_{bv}\alpha};\ \frac{A_{bu}}{3}\alpha;\ \frac{8\alpha\left(A_{bv}+A_{bu}/3\right)}{A_{\mathcal{K}}^2}\right\}\ . \tag{8.42}$$

Since by (**Abu**) we have

$$\left\|Ks_M - Ks_*\right\|_\infty \leq A_{bu}$$

and by (**Abv**) we have

$$\mathrm{Var}\left(Ks_M - Ks_*\right) \leq A_{bv} \times \ell\left(s_*, s_M\right)\ ,$$

we apply Bernstein's inequality (7.46) to $\bar{\delta}\left(M\right) = \left(P_n - P\right)\left(Ks_M - Ks_*\right)$ and we get for all $x > 0$,

$$\mathbb{P}\left(\left|\bar{\delta}\left(M\right)\right| \geq \sqrt{\frac{2A_{bv}\ell\left(s_*, s_M\right)x}{n}} + \frac{A_{bu}x}{3n}\right) \leq 2\exp\left(-x\right)\ .$$

By taking $x = \alpha\ln n$, we then have

$$\mathbb{P}\left(\left|\bar{\delta}\left(M\right)\right| \geq \sqrt{\frac{2A_{bv}\alpha\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{A_{bu}\alpha\ln n}{3n}\right) \leq 2n^{-\alpha}\ , \tag{8.43}$$

which gives the first part of Lemma 8.2 for A_d given in (8.42). Now, by noticing the fact that $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$, and by using it in (8.43) with $a = \ell\left(s_*, s_M\right)$, $b = \frac{A_{bv}\alpha\ln n}{n}$ and $\eta = D_M^{-1/2}$, we obtain

$$\mathbb{P}\left(\left|\bar{\delta}\left(M\right)\right| \geq \frac{\ell\left(s_*, s_M\right)}{\sqrt{D_M}} + \left(A_{bv}\sqrt{D_M} + \frac{A_{bu}}{3}\right)\frac{\alpha\ln n}{n}\right) \leq 2n^{-\alpha}\ . \tag{8.44}$$

Then, for a model $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}\left(\ln n\right)^2 \leq D_M$, we apply Lemma 8.1 and by (8.32), it holds for all $n \geq n_0\left((\mathbf{SA})\right)$,

$$\left(1 - 2\varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[p_2\left(M\right)\right] \tag{8.45}$$

where $\varepsilon_n = A_0\max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4},\ \left(\frac{D_M\ln n}{n}\right)^{1/4},\ \sqrt{R_{n,D_M,\alpha}}\right\}$. Moreover as $D_M \leq A_{\mathcal{M},+}n\left(\ln n\right)^{-2}$ by (**P2**), $R_{n,D_M} \leq A_{cons}\left(\ln n\right)^{-1/2}$ by (8.15) and $A_{\mathcal{M},+}\left(\ln n\right)^2 \leq D_M$, we deduce that for all $n \geq n_0\left(A_{\mathcal{M},+}, A_{cons}, A_0\right)$,

$$\varepsilon_n^2\left(M\right) \leq \frac{1}{4}\ .$$

Now, since $\mathcal{K}_{1,M} \geq A_{\mathcal{K}} > 0$ by (**AKl**), we have by (8.45), $\mathbb{E}\left[p_2\left(M\right)\right] \geq \frac{A_{\mathcal{K}}^2}{8}\frac{D_M}{n}$ for all $n \geq n_0\left((\mathbf{SA})\right)$. This allows, using (8.44), to conclude the proof by simple computations, for the value of A_d given in (8.42). ■

We turn now to the proofs of Theorems 4.2 and 4.1. These proofs follow from straightforward adaptations of the proofs of Theorems 4.2 and 4.1 given in Chapter 4.

Proof of Theorem 8.2. From the definition of the selected model $\widehat{M}$ given in (8.11), $\widehat{M}$ minimizes

$$\mathrm{crit}\left(M\right) := P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}\left(M\right)\ , \tag{8.46}$$

over the models $M \in \mathcal{M}_n$. Hence, $\widehat{M}$ also minimizes

$$\mathrm{crit}'\left(M\right) := \mathrm{crit}\left(M\right) - P_n\left(Ks_*\right)\ . \tag{8.47}$$

over the collection $\mathcal{M}_n$. Let us write

$$\begin{aligned}\ell\left(s_*, s_n\left(M\right)\right) &= P\left(Ks_n\left(M\right) - Ks_*\right) \\ &= P_n\left(Ks_n\left(M\right)\right) + P_n\left(Ks_M - Ks_n\left(M\right)\right) + \left(P_n - P\right)\left(Ks_* - Ks_M\right) \\ &+ P\left(Ks_n\left(M\right) - Ks_M\right) - P_n\left(Ks_*\right) \ .\end{aligned}$$

By setting

$$\begin{aligned}p_1\left(M\right) &= P\left(Ks_n\left(M\right) - Ks_M\right) \ , \\ p_2\left(M\right) &= P_n\left(Ks_M - Ks_n\left(M\right)\right) \ , \\ \bar{\delta}\left(M\right) &= \left(P_n - P\right)\left(Ks_M - Ks_*\right)\end{aligned}$$

and

$$\mathrm{pen}'_{\mathrm{id}}\left(M\right) = p_1\left(M\right) + p_2\left(M\right) - \bar{\delta}\left(M\right) \ ,$$

we have

$$\ell\left(s_*, s_n\left(M\right)\right) = P_n\left(Ks_n\left(M\right)\right) + p_1\left(M\right) + p_2\left(M\right) - \bar{\delta}\left(M\right) - P_n\left(Ks_*\right) \tag{8.48}$$

and by (8.47),

$$\mathrm{crit}'\left(M\right) = \ell\left(s_*, s_n\left(M\right)\right) + \left(\mathrm{pen}\left(M\right) - \mathrm{pen}'_{\mathrm{id}}\left(M\right)\right) \ . \tag{8.49}$$

As $\widehat{M}$ minimizes crit' over $\mathcal{M}_n$, it is therefore sufficient by (8.49), to control $\mathrm{pen}\left(M\right) - \mathrm{pen}'_{\mathrm{id}}\left(M\right)$ - or equivalently $\mathrm{crit}'\left(M\right)$ - in terms of the excess risk $\ell\left(s_*, s_n\left(M\right)\right)$, for every $M \in \mathcal{M}_n$, in order to derive oracle inequalities. Let Ω_n be the event on which:

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_M$, (8.21) hold and

$$\left|p_1\left(M\right) - \mathbb{E}\left[p_2\left(M\right)\right]\right| \leq L_{(\mathbf{SA})}\varepsilon_n\left(M\right)\mathbb{E}\left[p_2\left(M\right)\right] \tag{8.50}$$

$$\left|p_2\left(M\right) - \mathbb{E}\left[p_2\left(M\right)\right]\right| \leq L_{(\mathbf{SA})}\varepsilon_n^2\left(M\right)\mathbb{E}\left[p_2\left(M\right)\right] \tag{8.51}$$

$$\left|\bar{\delta}\left(M\right)\right| \leq \frac{\ell\left(s_*, s_M\right)}{\sqrt{D_M}} + L_{(\mathbf{SA})}\frac{\ln n}{\sqrt{D_M}}\mathbb{E}\left[p_2\left(M\right)\right] \tag{8.52}$$

$$\left|\bar{\delta}\left(M\right)\right| \leq L_{(\mathbf{SA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{8.53}$$

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $D_M \leq A_{\mathcal{M},+}\left(\ln n\right)^3$, (8.22) holds together with

$$\left|\bar{\delta}\left(M\right)\right| \leq L_{(\mathbf{SA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{8.54}$$

$$p_2\left(M\right) \leq L_{(\mathbf{SA})}\frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{SA})}\frac{\left(\ln n\right)^3}{n} \tag{8.55}$$

$$p_1\left(M\right) \leq L_{(\mathbf{SA})}\frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{SA})}\frac{\left(\ln n\right)^3}{n} \tag{8.56}$$

By (8.28), (8.29), (8.30) and (8.31) in Remark 8.1, Lemma 8.1, Lemma 8.2 applied with $\alpha = 2 + \alpha_{\mathcal{M}}$, and since (8.21) holds with probability at least $1 - A_p n^{-2}$, we get for all $n \geq n_0\left((\mathbf{SA})\right)$,

$$\mathbb{P}\left(\Omega_n\right) \geq 1 - A_p n^{-2} - 24\sum_{M \in \mathcal{M}_n} n^{-2-\alpha_{\mathcal{M}}} \geq 1 - L_{A_p, c_{\mathcal{M}}} n^{-2} \ .$$

Control on the criterion crit' for models of dimension not too small:

We consider models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$. Notice that (8.52) implies by (8.27) that, for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$, for all $n \geq n_0 ((\mathbf{SA}))$,

$$\begin{aligned} \left|\bar{\delta}(M)\right| &\leq L_{(\mathbf{SA})} \left(\frac{(\ln n)^3}{D_M} \cdot \frac{\ln n}{D_M} \right)^{1/4} \times \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \\ &\leq L_{(\mathbf{SA})} \varepsilon_n(M) \, \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] , \end{aligned}$$

so that on Ω_n we have, for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$,

$$\begin{aligned} &\left|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)\right| \\ &\leq \left|p_1(M) + p_2(M) - \mathrm{pen}(M)\right| + \left|\bar{\delta}(M)\right| \\ &\leq \left|p_1(M) + p_2(M) - 2\mathbb{E}\left[p_2(M)\right]\right| + \delta \mathbb{E}\left[p_2(M)\right] + L_{(\mathbf{SA})} \varepsilon_n(M) \, \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \\ &\leq L_{(\mathbf{SA})} \varepsilon_n(M) \, \mathbb{E}\left[p_2(M)\right] + \delta \mathbb{E}\left[p_2(M)\right] + L_{(\mathbf{SA})} \varepsilon_n(M) \, \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \\ &\leq \left(\delta + L_{(\mathbf{SA})} \varepsilon_n(M)\right) \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] . \end{aligned} \tag{8.57}$$

Now notice that using (**P2**) and (8.15) in (8.27) gives that for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0 ((\mathbf{SA}))$, $0 < L_{(\mathbf{SA})} \varepsilon_n(M) \leq \frac{1}{2}$. As $\ell(s_*, s_n(M)) = \ell(s_*, s_M) + p_1(M)$, we thus have on Ω_n, for all $n \geq n_0 ((\mathbf{SA}))$,

$$\begin{aligned} 0 &\leq \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \\ &\leq \ell(s_*, s_n(M)) + \left|p_1(M) - \mathbb{E}\left[p_2(M)\right]\right| \\ &\leq \ell(s_*, s_n(M)) + \frac{L_{(\mathbf{SA})} \varepsilon_n(M)}{1 - L_{(\mathbf{SA})} \varepsilon_n(M)} p_1(M) \qquad \text{by (8.50)} \\ &\leq \frac{1 + L_{(\mathbf{SA})} \varepsilon_n(M)}{1 - L_{(\mathbf{SA})} \varepsilon_n(M)} \ell(s_*, s_n(M)) \\ &\leq \left(1 + L_{(\mathbf{SA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) . \end{aligned} \tag{8.58}$$

Hence, using (8.58) in (8.57), we have on Ω_n for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0 ((\mathbf{SA}))$,

$$\left|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)\right| \leq \left(\delta + L_{(\mathbf{SA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) . \tag{8.59}$$

By consequence, for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0 ((\mathbf{SA}))$, it holds on Ω_n, using (8.49) and (8.59),

$$\left(1 - \delta - L_{(\mathbf{SA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) \leq \mathrm{crit}'(M) \leq \left(1 + \delta + L_{(\mathbf{SA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) . \tag{8.60}$$

Control on the criterion crit′ **for models of small dimension:**

We consider models $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$. By (8.22), (8.54) and (8.55), it holds on Ω_n, for any $\tau > 0$ and for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

$$\begin{aligned} &\left|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)\right| \\ &\leq p_1(M) + p_2(M) + \mathrm{pen}(M) + \left|\bar{\delta}(M)\right| \\ &\leq L_{(\mathbf{SA})} \frac{(\ln n)^3}{n} + A_r \frac{(\ln n)^3}{n} + L_{(\mathbf{SA})} \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right) \\ &\leq L_{(\mathbf{SA}), A_r} \frac{(\ln n)^3}{n} + \tau \ell(s_*, s_M) + \left(\tau^{-1} + 1\right) L_{(\mathbf{SA})} \frac{\ln n}{n} \\ &\leq L_{(\mathbf{SA}), A_r} \frac{(\ln n)^3}{n} + \tau \ell(s_*, s_n(M)) + \left(\tau^{-1} + 1\right) L_{(\mathbf{SA})} \frac{\ln n}{n} . \end{aligned} \tag{8.61}$$

Hence, by taking $\tau = (\ln n)^{-2}$ in (8.61) we get that for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$, it holds on Ω_n,

$$\left|\mathrm{pen}'_{\mathrm{id}} (M) - \mathrm{pen} (M)\right| \leq \frac{\ell (s_*, s_n (M))}{(\ln n)^2} + L_{(\mathbf{SA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{8.62}$$

Moreover, by (8.49) and (8.62), we have on the event Ω_n, for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

$$\left(1 - (\ln n)^{-2}\right) \ell (s_*, s_n (M)) - L_{(\mathbf{SA}),A_r} \frac{(\ln n)^3}{n} \leq \mathrm{crit}' (M) \tag{8.63}$$

$$\leq \left(1 + (\ln n)^{-2}\right) \ell (s_*, s_n (M)) + L_{(\mathbf{SA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{8.64}$$

Oracle inequalities:

Recall that by the definition given in (8.10), an oracle model satisfies

$$M_* \in \arg \min_{M \in \mathcal{M}_n} \{\ell (s_*, s_n (M))\} \ . \tag{8.65}$$

By Lemmas 8.3 and 8.4 below, we control on Ω_n the dimensions of the selected model $\widehat{M}$ and the oracle model M_*. More precisely, by (8.77) and (8.79), we have on Ω_n, for any $\frac{1}{2} > \eta > (1 - \beta_+)_+ /2$ and for all $n \geq n_0 ((\mathbf{SA}), \eta, \delta)$,

$$D_{\widehat{M}} \leq n^{1/2+\eta} \ , \tag{8.66}$$

$$D_{M_*} \leq n^{1/2+\eta} \ . \tag{8.67}$$

Now, from (8.66) we distinguish two cases in order to control $\mathrm{crit}' \left(\widehat{M}\right)$. If $A_{\mathcal{M},+} (\ln n)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta}$, we get by (8.60), for all $n \geq n_0 ((\mathbf{SA}))$,

$$\mathrm{crit}' \left(\widehat{M}\right) \geq \left(1 - \delta - L_{(\mathbf{SA})} \varepsilon_n \left(\widehat{M}\right)\right) \ell \left(s_*, s_n \left(\widehat{M}\right)\right) \ . \tag{8.68}$$

Otherwise, if $D_{\widehat{M}} \leq A_{\mathcal{M},+} (\ln n)^3$, we get by (8.63),

$$\left(1 - (\ln n)^{-2}\right) \ell \left(s_*, s_n \left(\widehat{M}\right)\right) - L_{(\mathbf{SA}),A_r} \frac{(\ln n)^3}{n} \leq \mathrm{crit}' \left(\widehat{M}\right) \ . \tag{8.69}$$

In all cases, we have by (8.68) and (8.69), for all $n \geq n_0 ((\mathbf{SA}))$,

$$\mathrm{crit}' \left(\widehat{M}\right) \geq \left(1 - \delta - (\ln n)^{-2} - L_{(\mathbf{SA})} \sup_{M \in \mathcal{M}_n,\ A_{\mathcal{M},+} (\ln n)^3 \leq D_M \leq n^{1/2+\eta}} \varepsilon_n (M)\right) \ell \left(s_*, s_n \left(\widehat{M}\right)\right)$$

$$- L_{(\mathbf{SA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{8.70}$$

Similarly, from (8.67) we distinguish two cases in order to control $\mathrm{crit}' (M_*)$. If $A_{\mathcal{M},+} (\ln n)^3 \leq D_{M_*} \leq n^{1/2+\eta}$, we get by (8.60), for all $n \geq n_0 ((\mathbf{SA}))$,

$$\mathrm{crit}' (M_*) \leq \left(1 + \delta + L_{(\mathbf{SA})} \varepsilon_n (M_*)\right) \ell (s_*, s_n (M_*)) \ . \tag{8.71}$$

Otherwise, if $D_{M_*} \leq A_{\mathcal{M},+} (\ln n)^3$, we get by (8.64),

$$\mathrm{crit}' (M_*) \leq \left(1 + (\ln n)^{-2}\right) \ell (s_*, s_n (M_*)) + L_{(\mathbf{SA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{8.72}$$

In all cases, we deduce from (8.71) and (8.72) that we have for all $n \geq n_0((\mathbf{SA}),\delta)$,

$$\mathrm{crit}'(M_*) \leq \left(1+\delta+(\ln n)^{-2}+L_{(\mathbf{SA})} \sup_{M\in\mathcal{M}_n,\ A_{\mathcal{M},+}(\ln n)^3\leq D_M\leq n^{1/2+\eta}} \varepsilon_n(M)\right) \ell(s_*, s_n(M_*)) + L_{(\mathbf{SA}),A_r}\frac{(\ln n)^3}{n} \ . \tag{8.73}$$

Hence, by setting

$$\theta_n = L_{(\mathbf{SA})} \times \sup_{M\in\mathcal{M}_n,\ A_{\mathcal{M},+}(\ln n)^3\leq D_M\leq n^{1/2+\eta}} \varepsilon_n(M) \ ,$$

we have by (8.27) and (8.15), for all $n \geq n_0((\mathbf{SA}),\eta,\delta)$,

$$\theta_n \leq \frac{L_{(\mathbf{SA})}}{(\ln n)^{1/4}} \ , \quad (\ln n)^{-2}+\theta_n+\delta<1 \ , \quad (\ln n)^{-2}+\theta_n<\frac{1-\delta}{2}$$

and we deduce from (8.70) and (8.73), since $\frac{1}{1-x}\leq 1+2x$ for all $x\in\left[0,\frac{1}{2}\right)$, that for all $n\geq n_0((\mathbf{SA}),\eta,\delta)$, it holds on Ω_n,

$$\begin{aligned}\ell\left(s_*, s_n\left(\widehat{M}\right)\right) &\leq \left(\frac{1+\delta+(\ln n)^{-2}+\theta_n}{1-\delta-(\ln n)^{-2}-\theta_n}\right)\ell(s_*,s_n(M_*)) + \frac{L_{(\mathbf{SA}),A_r}}{1-\delta-(\ln n)^{-2}-\theta_n}\frac{(\ln n)^3}{n} \\ &\leq \left(\frac{1+\delta}{1-\delta}+\frac{5\left((\ln n)^{-2}+\theta_n\right)}{(1-\delta)^2}\right)\ell(s_*,s_n(M_*)) + L_{(\mathbf{SA}),A_r}\frac{(\ln n)^3}{n} \ . \end{aligned} \tag{8.74}$$

Inequality (8.24) is now proved.
It remains to prove the second part of Theorem 8.2. We assume that assumption (**Ap**) holds. From Lemmas 8.3 and 8.4, we have that for any $\frac{1}{2}>\eta>(1-\beta_+)_+/2$ and for all $n\geq n_0((\mathbf{SA}),C_-,\beta_-,\eta,\delta)$, it holds on Ω_n,

$$A_{\mathcal{M},+}(\ln n)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta} \ , \tag{8.75}$$

$$A_{\mathcal{M},+}(\ln n)^3 \leq D_{M_*} \leq n^{1/2+\eta} \ . \tag{8.76}$$

Now, using (8.68) and (8.71), by the same kind of computations leading to (8.74), we deduce that it holds on Ω_n, for all $n\geq n_0((\mathbf{SA}),C_-,\beta_-,\eta,\delta)$,

$$\begin{aligned}\ell\left(s_*, s_n\left(\widehat{M}\right)\right) &\leq \left(\frac{1+\delta+\theta_n}{1-\delta-\theta_n}\right)\ell(s_*,s_n(M_*)) \\ &\leq \left(\frac{1+\delta}{1-\delta}+\frac{5\theta_n}{(1-\delta)^2}\right)\ell(s_*,s_n(M_*)) \ . \end{aligned}$$

Thus inequality (8.25) is proved and Theorem 8.2 follows. ■

Lemma 8.3 (Control on the dimension of the selected model) *Assume that the general set of assumptions (**SA**) hold. Let $\eta>(1-\beta_+)_+/2$. If $n\geq n_0((\mathbf{SA}),\eta,\delta)$ then, on the event Ω_n defined in the proof of Theorem 8.2, it holds*

$$D_{\widehat{M}} \leq n^{1/2+\eta} \ . \tag{8.77}$$

*If moreover (**Ap**) holds, then for all $n\geq n_0((\mathbf{SA}),C_-,\beta_-,\eta,\delta)$, we have on the event Ω_n,*

$$A_{\mathcal{M},+}(\ln n)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta} \ . \tag{8.78}$$

Lemma 8.4 (Control on the dimension of oracle models) *Assume that the general set of assumptions (**SA**) hold. Let $\eta > (1-\beta_+)_+/2$. If $n \geq n_0((\boldsymbol{SA}),\eta)$ then, on the event Ω_n defined in the proof of Theorem 8.2, it holds*

$$D_{M_*} \leq n^{1/2+\eta} \ . \tag{8.79}$$

*If moreover (**Ap**) holds, then for all $n \geq n_0((\boldsymbol{SA}), C_-, \beta_-, \eta)$, we have on the event Ω_n,*

$$A_{\mathcal{M},+} (\ln n)^3 \leq D_{M_*} \leq n^{1/2+\eta} \ . \tag{8.80}$$

Proof of Lemma 8.3. Recall that $\widehat{M}$ minimizes

$$\mathrm{crit}'(M) = \mathrm{crit}(M) - P_n K s_* = \ell(s_*, s_M) - p_2(M) + \bar{\delta}(M) + \mathrm{pen}(M) \tag{8.81}$$

over the models $M \in \mathcal{M}_n$.

1. Lower bound on $\mathrm{crit}'(M)$ for small models in the case where (**Ap**) hold : let $M \in \mathcal{M}_n$ be such that $D_M < A_{\mathcal{M},+} (\ln n)^3$. We then have on Ω_n,

$$\begin{aligned}
&\ell(s_*, s_M) \geq C_- A_{\mathcal{M},+}^{-\beta_-} (\ln n)^{-3\beta_-} \text{ by (Ap)} \\
&\mathrm{pen}(M) \geq 0 \\
&p_2(M) \leq L_{(\mathbf{SA})} \frac{(\ln n)^3}{n} \qquad \text{from (8.55)} \\
&\bar{\delta}(M) \geq -L_{(\mathbf{SA})} \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right) \text{ from (8.54).}
\end{aligned}$$

   Since by (**Abu**), we have $0 \leq \ell(s_*, s_M) \leq \|K s_M - K s_*\|_\infty \leq A_{bu}$, we deduce that for all $n \geq n_0((\mathbf{SA}), C_-, \beta_-)$,

$$\mathrm{crit}'(M) \geq \frac{C_- A_{\mathcal{M},+}^{-\beta_-}}{2} (\ln n)^{-3\beta_-} \ . \tag{8.82}$$

2. Lower bound for large models : let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/2+\eta}$. From (8.21) and (8.51) we have on Ω_n,

$$\mathrm{pen}(M) - p_2(M) \geq \left(1 - \delta - L_{(\mathbf{SA})} \varepsilon_n^2(M)\right) \mathbb{E}\left[p_2(M)\right] \ .$$

   Using (**P2**), (8.15) and the fact that $D_M \geq n^{1/2+\eta}$ in (8.27), we deduce that for all $n \geq n_0((\mathbf{SA}), \eta, \delta)$, $L_{(\mathbf{SA})} \varepsilon_n^2(M) \leq \frac{1}{2}(1-\delta)$ and as by (**AKl**), $\mathcal{K}_{1,M} \geq A_{\mathcal{K}} > 0$ we also deduce from Lemma 8.1 that for all $n \geq n_0((\mathbf{SA}), \eta)$, $\mathbb{E}\left[p_2(M)\right] \geq \frac{A_{\mathcal{K}}^2}{8} \frac{D_M}{n}$. By consequence, it holds for all $n \geq n_0((\mathbf{SA}), \eta, \delta)$,

$$\mathrm{pen}(M) - p_2(M) \geq \frac{A_{\mathcal{K}}^2}{16} (1-\delta) \frac{D_M}{n} \ . \tag{8.83}$$

   From (8.53) it holds on Ω_n,

$$\bar{\delta}(M) \geq -L_{(\mathbf{SA})} \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right) \ . \tag{8.84}$$

   Hence, as $D_M \geq n^{1/2+\eta}$ and as by (**Abu**), $0 \leq \ell(s_*, s_M) \leq A_{bu}$, we deduce from (8.81), (8.83) and (8.84) that we have on Ω_n, for all $n \geq n_0((\mathbf{SA}), \eta, \delta)$,

$$\mathrm{crit}'(M) \geq (1-\delta) L_{(\mathbf{SA})} n^{-1/2+\eta} \ . \tag{8.85}$$

3. A better model exists for $\mathrm{crit}'(M)$: from (**P3**), there exists $M_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n}$. Then, for all $n \geq n_0((\mathbf{SA}), \eta)$,

$$A_{\mathcal{M},+}(\ln n)^3 \leq \sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n} \leq n^{1/2+\eta} .$$

Using ($\mathbf{Ap}_u$),

$$\ell(s_*, s_{M_0}) \leq C_+ n^{-\beta_+/2} . \tag{8.86}$$

By (8.52), we have on Ω_n, for all $n \geq n_0((\mathbf{SA}), \eta)$,

$$\left|\bar{\delta}(M_0)\right| \leq \frac{\ell(s_*, s_{M_0})}{\sqrt{D_{M_0}}} + L_{(\mathbf{SA})}\frac{\ln n}{\sqrt{D_{M_0}}}\mathbb{E}[p_2(M_0)] \tag{8.87}$$

and by (8.21),

$$\mathrm{pen}(M_0) \leq 3\mathbb{E}[p_2(M_0)] .$$

Hence, as by (8.12) and (7.23) we have $\mathcal{K}_{1,M} \leq A_1 A_H$ and by (**Abu**) $\ell(s_*, s_{M_0}) \leq A_{bu}$ by (**Ab**) and as for all $n \geq n_0((\mathbf{SA}))$ $\varepsilon_n(M) \leq 1$, we deduce from inequalities (8.86), (8.87) and Lemma 8.1 that for all $n \geq n_0((\mathbf{SA}), \eta)$,

$$\left|\bar{\delta}(M_0)\right| \leq L_{(\mathbf{SA})}\left(n^{-(\beta_+/2+1/4)} + \ln(n)\, n^{-3/4}\right)$$

and

$$\mathrm{pen}(M_0) \leq L_{(\mathbf{SA})} n^{-1/2} .$$

By consequence, we have on Ω_n, for all $n \geq n_0((\mathbf{SA}), \eta)$,

$$\begin{aligned}\mathrm{crit}'(M_0) &\leq \ell(s_*, s_{M_0}) + \left|\bar{\delta}(M_0)\right| + \mathrm{pen}(M_0) \\ &\leq L_{(\mathbf{SA})}\left(n^{-\beta_+/2} + n^{-1/2}\right) .\end{aligned} \tag{8.88}$$

To conclude, notice that the upper bound (8.88) is smaller than the lower bound given in (8.85) for all $n \geq n_0((\mathbf{SA}), \eta, \delta)$. Hence, points 2 and 3 above yield inequality (8.77). Moreover, the upper bound (8.88) is smaller than lower bounds given in (8.82), derived by using (**Ap**), and (8.85), for all $n \geq n_0((\mathbf{SA}), C_-, \beta_-, \eta, \delta)$. This thus gives (8.78) and Lemma 8.3 is proved. ■

Proof of Lemma 8.4. By definition, M_* minimizes

$$\ell(s_*, s_n(M)) = \ell(s_*, s_M) + p_1(M)$$

over the models $M \in \mathcal{M}_n$.

1. Lower bound on $\ell(s_*, s_n(M))$ for small models : let $M \in \mathcal{M}_n$ be such that $D_M < A_{\mathcal{M},+}(\ln n)^3$. In this case we have

$$\ell(s_*, s_n(M)) \geq \ell(s_*, s_M) \geq C_- A_{\mathcal{M},+}^{-\beta_-}(\ln n)^{-3\beta_-} \text{ by } (\mathbf{Ap}). \tag{8.89}$$

2. Lower bound of $\ell(s_*, s_n(M))$ for large models : let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/2+\eta}$. From (8.50) we get on Ω_n,

$$p_1(M) \geq \left(1 - L_{(\mathbf{SA})}\varepsilon_n(M)\right)\mathbb{E}[p_2(M)] .$$

Using (**P2**), (8.15) and the fact that $D_M \geq n^{1/2+\eta}$ in (8.27), we deduce that for all $n \geq n_0((\mathbf{SA}), \eta)$, $L_{(\mathbf{SA})}\varepsilon_n(M) \leq \frac{1}{2}$ and as by (**AKl**), $\mathcal{K}_{1,M} \geq A_{\mathcal{K}} > 0$ we also deduce from Lemma 8.1 that for all $n \geq n_0((\mathbf{SA}), \eta)$, $\mathbb{E}[p_2(M)] \geq \frac{A_{\mathcal{K}}^2}{8}\frac{D_M}{n}$. By consequence, it holds for all $n \geq n_0((\mathbf{SA}), \eta)$, on the event Ω_n,

$$\ell(s_*, s_n(M)) \geq p_1(M) \geq \frac{A_{\mathcal{K}}^2}{16}\frac{D_M}{n} \geq \frac{A_{\mathcal{K}}^2}{16} n^{-1/2+\eta} . \tag{8.90}$$

3. A better model exists for $\ell(s_*, s_n(M))$: from (**P3**), there exists $M_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n}$. Moreover, for all $n \geq n_0((\mathbf{SA}), \eta)$,

$$A_{\mathcal{M},+}(\ln n)^3 \leq \sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n} \leq n^{1/2+\eta} .$$

Using $(\mathbf{Ap}_u)$,

$$\ell(s_*, s_{M_0}) \leq C_+ n^{-\beta_+/2}$$

and by (8.50)

$$p_1(M_0) \leq \left(1 + L_{(\mathbf{SA})}\varepsilon_n(M)\right)\mathbb{E}[p_2(M_0)]$$

Hence, as by (8.12) and (7.23) we have $\mathcal{K}_{1,M} \leq A_1 A_H$ and as, by (8.15) and (8.27), for all $n \geq n_0((\mathbf{SA}))$ it holds $\varepsilon_n(M) \leq 1$, we deduce from Lemma 8.1 that for all $n \geq n_0((\mathbf{SA}))$, on the event Ω_n,

$$p_1(M_0) \leq L_{(\mathbf{SA})}\frac{D_M}{n} \leq L_{(\mathbf{SA})}n^{-1/2} .$$

By consequence, on Ω_n, for all $n \geq n_0((\mathbf{SA}))$,

$$\begin{aligned}\ell(s_*, s_n(M_0)) &= \ell(s_*, s_{M_0}) + p_1(M_0) \\ &\leq L_{(\mathbf{SA})}\left(n^{-\beta_+/2} + n^{-1/2}\right) . \end{aligned} \tag{8.91}$$

The upper bound (8.91) is smaller than the lower bound (8.90) for all $n \geq n_0((\mathbf{SA}), \eta)$, and this gives (8.79). If (**Ap**) hold, then the upper bound (8.91) is smaller than the lower bounds (8.89) and (8.90) for all $n \geq n_0((\mathbf{SA}), C_-, \beta_-, \eta)$, which proves (8.80) and allows to conclude the proof of Lemma 8.4. ■

Proof of Theorem 8.1. Similarly to the proof of Theorem 8.2, we consider the event Ω'_n of probability at least $1 - L_{c_{\mathcal{M}}, A_p}n^{-2}$ for all $n \geq n_0((\mathbf{SA}))$, on which: (8.19) holds and

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $A_{\mathcal{M},+}(\ln n)^2 \leq D_M$ it holds

$$|p_1(M) - \mathbb{E}[p_2(M)]| \leq L_{(\mathbf{SA})}\varepsilon_n(M)\mathbb{E}[p_2(M)] , \tag{8.92}$$

$$|p_2(M) - \mathbb{E}[p_2(M)]| \leq L_{(\mathbf{SA})}\varepsilon_n^2(M)\mathbb{E}[p_2(M)] . \tag{8.93}$$

- For all models $M \in \mathcal{M}_n$ with $D_M \leq A_{\mathcal{M},+}(\ln n)^2$ it holds

$$p_2(M) \leq L_{(\mathbf{SA})}\frac{(\ln n)^2}{n} . \tag{8.94}$$

- For every $M \in \mathcal{M}_n$,

$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{SA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) . \tag{8.95}$$

Let $d \in (0, 1)$ to be chosen later.

Lower bound on $D_{\widehat{M}}$. Remind that $\widehat{M}$ minimizes

$$\mathrm{crit}'(M) = \mathrm{crit}(M) - P_n K s_* = \ell(s_*, s_M) - p_2(M) + \bar{\delta}(M) + \mathrm{pen}(M) . \tag{8.96}$$

1. Lower bound on $\mathrm{crit}'(M)$ for "small" models : assume that $M \in \mathcal{M}_n$ and

$$D_M \leq dA_{rich} n (\ln n)^{-2} \ .$$

We have

$$\ell(s_*, s_M) + \mathrm{pen}(M) \geq 0 \tag{8.97}$$

and from (8.95), as $\ell(s_*, s_M) \leq A_{bu}$ by (**Abu**), we get on Ω'_n, for all $n \geq n_0((\mathbf{SA}),d)$,

$$\begin{aligned}\bar{\delta}(M) &\geq -L_{(\mathbf{SA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) \\ &\geq -L_{(\mathbf{SA})}\sqrt{\frac{\ln n}{n}} \\ &\geq -d \times (A_1 A_H)^2 A_{rich} (\ln n)^{-2} \ . \end{aligned} \tag{8.98}$$

Then, if $D_M \geq A_{\mathcal{M},+}(\ln n)^2$, as $\mathcal{K}_{1,M} \leq A_1 A_H$ by (8.12) and (7.23) and as, by (8.15) and (8.27), for all $n \geq n_0((\mathbf{SA}))$ it holds $L_{(\mathbf{SA})}\varepsilon_n(M) \leq 1$, we deduce from (8.93) and Lemma 8.1 that for all $n \geq n_0((\mathbf{SA}))$,

$$p_2(M) \leq 2\mathbb{E}[p_2(M)] \leq (A_1 A_H)^2 \frac{D_M}{n} \leq d \times (A_1 A_H)^2 A_{rich} (\ln n)^{-2} \ .$$

Whenever $D_M \leq A_{\mathcal{M},+}(\ln n)^2$, (8.94) gives that, for all $n \geq n_0((\mathbf{SA}),d)$, on the event Ω'_n,

$$p_2(M) \leq L_{(\mathbf{SA})}\frac{(\ln n)^2}{n} \leq d \times (A_1 A_H)^2 A_{rich} (\ln n)^{-2} \ .$$

Hence, we have checked that for all $n \geq n_0((\mathbf{SA}), d)$, on the event Ω'_n,

$$-p_2(M) \geq -d \times (A_1 A_H)^2 A_{rich} (\ln n)^{-2} \ , \tag{8.99}$$

and finally, by using (8.97), (8.98) and (8.99) in (8.96), we deduce that on Ω'_n, for all $n \geq n_0((\mathbf{SA}), d)$,

$$\mathrm{crit}'(M) \geq -d \times 2(A_1 A_H)^2 A_{rich} (\ln n)^{-2} \ . \tag{8.100}$$

2. There exists a better model for $\mathrm{crit}'(M)$: By (**P3**), for all $n \geq n_0(A_{\mathcal{M},+}, A_{rich})$ a model $M_1 \in \mathcal{M}_n$ exists such that

$$A_{\mathcal{M},+}(\ln n)^2 \leq \frac{A_{rich} n}{(\ln n)^2} \leq D_{M_1} \ .$$

We then have on Ω'_n,

$$\begin{aligned} \ell(s_*, s_{M_1}) &\leq A_{rich}^{-\beta_+}(\ln n)^{2\beta_+} n^{-\beta_+} && \text{by } (\mathbf{Ap}_u) \\ p_2(M_1) &\geq \left(1 - L_{(\mathbf{SA})}\varepsilon_n^2(M_1)\right)\mathbb{E}[p_2(M_1)] && \text{by (8.93)} \\ \mathrm{pen}(M_1) &\leq A_{\mathrm{pen}}\mathbb{E}[p_2(M_1)] && \text{by (8.19)} \\ \left|\bar{\delta}(M_1)\right| &\leq L_{(\mathbf{SA})}\sqrt{\frac{\ln n}{n}} && \text{by (8.95) and } (\mathbf{Abu}) \end{aligned}$$

and therefore,

$$\mathrm{crit}'(M_1) \leq \left(-1 + A_{\mathrm{pen}} + L_{(\mathbf{SA})}\varepsilon_n^2(M_1)\right)\mathbb{E}[p_2(M_1)] + L_{(\mathbf{SA})}\sqrt{\frac{\ln n}{n}} + A_{rich}^{-\beta_+}\frac{(\ln n)^{2\beta_+}}{n^{\beta_+}} \ . \tag{8.101}$$

Hence, as $-1+A_{\rm pen}<0$, and as by (8.15), (8.27), (**AKl**) and Lemma 8.1 it holds for all $n \geq n_0\left((\mathbf{SA}), A_{\rm pen}\right)$

$$L_{(\mathbf{SA})}\varepsilon_n^2\left(M_1\right) \leq \frac{1-A_{\rm pen}}{2} \quad \text{and} \quad \mathbb{E}\left[p_2\left(M_1\right)\right] \geq \frac{A_{\mathcal{K}}^2}{8}\frac{D_M}{n} \geq \frac{A_{\mathcal{K}}^2 A_{rich}}{8}\left(\ln n\right)^{-2} \ ,$$

we deduce from (8.101) that on Ω_n', for all $n \geq n_0\left((\mathbf{SA}), A_{\rm pen}\right)$,

$$\mathrm{crit}'\left(M_1\right) \leq -\frac{1}{16}\left(1-A_{\rm pen}\right)A_{\mathcal{K}}^2 A_{rich}\left(\ln n\right)^{-2} \ . \tag{8.102}$$

Now, by taking

$$0 < d = \frac{1}{33}\left(1-A_{\rm pen}\right)\left(\frac{A_{\mathcal{K}}}{A_1 A_H}\right)^2 < 1 \tag{8.103}$$

and by comparing (8.100) and (8.102), we deduce that on Ω_n', for all $n \geq n_0\left((\mathbf{SA}), A_{\rm pen}\right)$, for all $M \in \mathcal{M}_n$ such that $D_M \leq dA_{rich}n\left(\ln n\right)^{-2}$,

$$\mathrm{crit}'\left(M_1\right) < \mathrm{crit}'\left(M\right)$$

and so

$$D_{\widehat{M}} > dA_{rich}n\left(\ln n\right)^{-2} \ . \tag{8.104}$$

Excess Risk of $s_n\left(\widehat{M}\right)$. We take d with the value given in (8.103). First notice that for all $n \geq n_0\left(A_{\mathcal{M},+}, A_{rich}, d\right)$, we have $dA_{rich}n\left(\ln n\right)^{-2} \geq A_{\mathcal{M},+}\left(\ln n\right)^2$. Hence, for all $M \in \mathcal{M}_n$ such that $D_M \geq dA_{rich}n\left(\ln n\right)^{-2}$, by (8.15), (8.27), (**P2**), (**An**) and Lemma 8.1, it holds on Ω_n' for all $n \geq n_0\left((\mathbf{SA}), A_{\rm pen}\right)$, using (8.92),

$$\ell\left(s_*, s_n\left(M\right)\right) \geq p_1\left(M\right) \geq \frac{A_{\mathcal{K}}^2}{8}\frac{D_M}{n} \geq \frac{dA_{\mathcal{K}}^2 A_{rich}}{8}\left(\ln n\right)^{-2} \ .$$

By (8.104), we thus get that on Ω_n', for all $n \geq n_0\left((\mathbf{SA}), A_{\rm pen}\right)$,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \frac{dA_{\mathcal{K}}^2 A_{rich}}{8}\left(\ln n\right)^{-2} \ . \tag{8.105}$$

Moreover, the model M_0 defined in (**P3**) satisfies, for all $n \geq n_0\left((\mathbf{SA})\right)$,

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq \sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n}$$

and so using ($\mathbf{Ap}_u$),

$$\ell\left(s_*, s_{M_0}\right) \leq C_+ n^{-\beta_+/2} \ .$$

In addition, by (8.50),

$$p_1\left(M\right) \leq \left(1+L_{(\mathbf{SA})}\varepsilon_n\left(M\right)\right)\mathbb{E}\left[p_2\left(M\right)\right] \ .$$

Hence, as $\mathcal{K}_{1,M} \leq A_1 A_H$ by (8.12) and (7.23) and as, by (8.15) and (8.27), for all $n \geq n_0\left((\mathbf{SA})\right)$ it holds $\varepsilon_n\left(M\right) \leq 1$, we deduce from Lemma 8.1 that for all $n \geq n_0\left((\mathbf{SA})\right)$,

$$p_1\left(M\right) \leq L_{(\mathbf{SA})}\frac{D_M}{n} \leq L_{(\mathbf{SA})}n^{-1/2} \ .$$

By consequence, for all $n \geq n_0\left((\mathbf{SA})\right)$,

$$\ell\left(s_*, s_n\left(M_0\right)\right) \leq L_{(\mathbf{SA})}\left(n^{-\beta_+/2}+n^{-1/2}\right) \tag{8.106}$$

and the ratio between the two bounds (8.105) and (8.106) is larger than $\ln\left(n\right)$ for all $n \geq n_0\left(L_{(\mathbf{SA})}, A_{\rm pen}\right)$, which yields (8.20). ■

Conclusions and prospects

En se rendant à Chartres, Péguy voit sur le bord de la route un homme qui casse des cailloux à grands coups de maillet. Son visage exprime le malheur et ses gestes la rage. Péguy s'arrête et demande : "Monsieur, que faites-vous ? " "Vous voyez bien, lui répond l'homme, je n'ai trouvé que ce métier stupide et douloureux. "Un peu plus loin, Péguy aperçoit un autre homme qui, lui aussi, casse des cailloux, mais son visage est calme et ses gestes harmonieux. "Que faites-vous, monsieur ? ", lui demande Péguy. "Eh bien, je gagne ma vie grâce à ce métier fatigant, mais qui a l'avantage d'être en plein air ", lui répond-il. Plus loin, un troisième casseur de cailloux irradie de bonheur. Il sourit en abattant la masse et regarde avec plaisir les éclats de pierre. "Que faites-vous ? ", lui demande Péguy. "Moi, répond cet homme, je bâtis une cathédrale ! "

BORIS CYRULNIK

Along this manuscript, which apart if it is explicitly mentioned, is a personal and original work of the author, we have developed a methodology based on the concept of regular contrast that allows to derive upper and lower bounds for the empirical and true excess risks on a fixed model that are optimal at the first order, and we have applied these results to validate the slope heuristics in several classical frameworks. We give now some prospects based on these advances.

Consistency of M-estimators in regular contrast estimation

A central issue in regular contrast estimation is the behavior in sup-norm of the considered M-estimators. More precisely, in order to control second order terms that appear in the expansion of a regular contrast, we require that the considered M-estimator for a fixed model is consistent in sup-norm towards the projection of the target onto the model, at a rate at least $(\ln n)^{-1/2}$. Considering models of dimension proportional to $n\,(\ln n)^{-2}$, the rate of convergence in sup-norm must be close to

$$\sqrt{\frac{D \ln n}{n}}\ .$$

We show that this is satisfied in particular cases, such as histograms and more general models of piecewise polynomials in the least-squares regression setting, or histogram models of densities when considering the maximum likelihood density estimation. An important problem in regular contrast estimation is thus to find some systematical way to derive the required consistency in sup-norm of the M-estimators. Based on the same type of ideas than those exposed in Section 7.3 of Chapter 7, we propose the following systematical approach, which for now, is a work in progress.

Consider a regular contrast $K : \mathcal{S} \longrightarrow L_1^-(P)$ in the sense of Definition 2.6 and let $M \subset \mathcal{S} \bigcap L_\infty(P)$ be a model. Let us set, for any $C \geq 0$, the slices of the model related to the sup-norm,

$$\mathcal{F}_C^\infty = \{s \in M \ ; \ \|s - s_M\|_\infty \leq C\} \ ,$$
$$\mathcal{F}_{>C}^\infty = \{s \in M \ ; \ \|s - s_M\|_\infty > C\} = (\mathcal{F}_C^\infty)^c \ ,$$
$$\mathcal{D}_C^\infty = \{s \in M \ ; \ \|s - s_M\|_\infty = C\} \ ,$$

and, for an interval $I \subset \mathbb{R}_+$,

$$\mathcal{F}_I^\infty = \{s \in M \ ; \ \|s - s_M\|_\infty \in I\} \ .$$

Then, for any $C \geq 0$, we write

$$\begin{aligned}
&\mathbb{P}\left(\|s_n(M) - s_M\|_\infty > C\right) \\
&\leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks - Ks_M) \leq \inf_{s \in \mathcal{F}_C^\infty} P_n(Ks - Ks_M)\right) \\
&= \mathbb{P}\left(\sup_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_C^\infty} P_n(Ks_M - Ks)\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\left(\|s_n(M) - s_M\|_\infty \leq C\right) \\
&\leq \mathbb{P}\left(\inf_{s \in \mathcal{F}_C^\infty} P_n(Ks - Ks_M) \leq \inf_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks - Ks_M)\right) \\
&= \mathbb{P}\left(\sup_{s \in \mathcal{F}_C^\infty} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_{>C}^\infty} P_n(Ks_M - Ks)\right) \\
&\leq \mathbb{P}\left(\sup_{s \in \mathcal{F}_C^\infty} P_n(Ks_M - Ks) \geq \sup_{s \in \mathcal{F}_{(C,rC]}^\infty} P_n(Ks_M - Ks)\right) \ ,
\end{aligned}$$

for any $r > 1$. Now, to fix ideas, take for instance K to be the least-squares regression contrast. We have for instance, for any $L \geq 0$,

$$\begin{aligned}
&\sup_{s \in \mathcal{D}_L^\infty} P_n(Ks_M - Ks) \\
&= \sup_{s \in \mathcal{D}_L^\infty} \left\{(P_n - P)(\psi_{1,M} \cdot (s_M - s)) - (P_n - P)(s_M - s)^2 - \|s - s_M\|_2^2\right\} \\
&\geq \sup_{s \in \mathcal{D}_L^\infty} (P_n - P)(\psi_{1,M} \cdot (s_M - s)) - \sup_{s \in \mathcal{D}_L^\infty} (P_n - P)(s_M - s)^2 - \sup_{s \in \mathcal{D}_L^\infty} \|s - s_M\|_2^2 \ .
\end{aligned}$$

In order to derive upper bounds and also lower bounds for the rates of convergence in sup-norm of the M-estimator, it remains to connect the sup-norm with the excess risk and the Hilbertian norm $\|\cdot\|_{H,M}$ which in the regression case are given by the quadratic norm $\|\cdot\|_2$. For instance, if M is a histogram model, given by

$$M = \left\{s = \sum_{I \in \Lambda_M} \beta_I \mathbf{1}_I \ ; \ (\beta_I)_{I \in \Lambda_M} \in \mathbb{R}^D\right\}$$

for a finite partition Λ_M, with $\mathrm{Card}(\Lambda_M) := D$, we can exactly compute the slices in terms of coordinates of their functions in the natural basis of M, made of the indicators of the elements of the partition Λ_M.

More precisely, we have

$$\mathcal{F}_C^\infty = \{s \in M \ ; \ \|s - s_M\|_\infty \leq C\} = \left\{ \sum_{I \in \Lambda_M} \beta_I \mathbf{1}_I \ ; \ \sup_{I \in \Lambda_M} |\beta_I| \leq C \right\} ,$$

$$\mathcal{F}_{>C}^\infty = \{s \in M \ ; \ \|s - s_M\|_\infty > C\} = \left\{ \sum_{I \in \Lambda_M} \beta_I \mathbf{1}_I \ ; \ \sup_{I \in \Lambda_M} |\beta_I| > C \right\} ,$$

and

$$\mathcal{D}_L^\infty = \{s \in M \ ; \ \|s - s_M\|_\infty = L\} = \left\{ \sum_{I \in \Lambda_M} \beta_I \mathbf{1}_I \ ; \ \sup_{I \in \Lambda_M} |\beta_I| = C \right\} .$$

To control the slices in sup-norm in terms of coordinates in an orthonormal basis associated to the Hilbertian norm $\|\cdot\|_{H,M}$, it seems that the localized basis assumption is a convenient property, and let us recall that in the histogram regression case this property is equivalent the lower-regularity of the partition Λ_M, which claims the existence of a positive constant $c_{M,P}$ such that

$$D \inf_{I \in \mathcal{P}} P^X (I) \geq c_{M,P} > 0 .$$

However, the localized basis property seems to be not sufficient for this systematical approach, which would more precisely require a control of the sup-norm of the functions in the model from above *and from below* in terms of their coordinates in a suitable basis.

More examples of regular contrasts on suitable models

The three examples of regular contrast estimation studied in this manuscript are classical non-parametric frameworks, and allow us to recover most of the recent results related to the theoretical study of the slope heuristics, initiated by Birgé and Massart in [23]. A central task is now to further investigate the scope of regular contrast estimation. For a M-estimation problem, existence of a projection of the considered target onto a model can be derived from arguments of compacity. Unicity of the projection typically follows from convexity arguments. Hence, we have now to find situations where both an expansion of the contrast and an equivalence of the excess risk on the model with a suitable Hilbertian norm can be derived. A careful look at the three examples of regular contrasts derived in Chapter 2 seems to indicate that the equivalence of an excess risk with some Hilbertian norm follows in particular from some kind of orthogonality of the model with respect to the target and more precisely that Pythagorean-like identities play a center role in this question.
Even if the binary classification setting stated in Section 2.1 of Chapter 2 seems, in its full generality, to be beyond the reach of our regular context, it nonetheless connected to convex frameworks employed in SVM, boosting or logistic regression methods, see for instance Bartlett & *al* [18]. These procedures are practically tractable by the use of a convex surrogate ϕ of the 0-1 loss in the minimization problem, that corresponds to a contrast of the type

$$K : s \longmapsto (Ks : z = (x, y) \longmapsto (Ks)(z) = \phi (y \cdot s(x))) \ . \tag{8.107}$$

Moreover, typical models in this context are taken to be scaled convex hull of a finite dimensional base class. If the convex function ϕ is smooth, it highly likely that the contrast stated in (8.107) could be expanded as required in the definition of a regular contrast, the convexity property as well as the scaling one allow to think that accurate risk bound could be derived in this setting. The proximity of the ϕ-risk with an Hilbertian norm has also to be tackled in order to strictly recover a regular contrast estimation setting, and consequently the possibility to derive accurate lower bounds of convergence, at least for the ϕ-risk.

Arlot's resampling penalties

Resampling penalties and V-fold penalties introduced by Arlot respectively in [5] and [7] are of general purpose and proved to be efficient in least-squares regression with random design and heteroscedastic noise, on histograms models. In particular, this framework is sufficiently general to show that Arlot's resampling and V-fold penalties give accurate data-driven penalization procedures, even in the case where the noise is highly heteroscedastic, a case where the ideal penalty is typically not a function of the dimension of the models and where linear penalties are - highly - sub-optimal, see Arlot [6].
For V-fold and sub-sampling cases, the control of Arlot's penalties can be derived for the concentration inequalities derived from the empirical and true risk. It is thus natural to think that such penalization methods are indeed efficient in general regular contrast estimation. A work is at its beginning on this subject with Sylvain Arlot, and sharp results seem to be attainable, combining Arlot's methods and ours, at least in the V-fold case.

Sparse recovery problems

It seems that the notion of regular contrast is closely related to the notion of "loss function of quadratic type" introduced in sparse recovery problems, see for instance Koltchinskii [47], [46], and see also Koltchinskii [45] for a study of the Dantzig selector in regression with random design. A challenging problem, which is at our top priorities, would be to adapt our methods to derive lower bounds for the rates of convergence of the excess risk in this context. It seems that exact constant would be only possible if the dictionary is assumed to be "$L_2\,(\Pi)$-orthonormal", where Π is a probability measure on the set defining the elements of the dictionary.

Bibliographie

[1] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[3] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. System identification and time-series analysis.

[4] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

[5] Sylvain Arlot. Model selection by resampling penalization, March 2008. oai:hal.archives-ouvertes.fr:hal-00262478_v2.

[6] Sylvain Arlot. Suboptimality of penalties proportional to the dimension for model selection in heteroscedastic regression, December 2008. arXiv:0812.3141.

[7] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

[8] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties, September 2009. arXiv:0909.1884v1.

[9] Sylvain Arlot and Peter L. Bartlett. Margin adaptive model selection in statistical learning, 2008. arXiv:0804.2937.

[10] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[11] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[12] A. R. Barron. Limits of information, markov chains, and projections. In *Proceedings. 2000 International Synopsium on Information Theory*, page 25, 2000.

[13] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[14] A.R. Barron. Approximation and estimation bounds for artificial neural networks. *Mach. Learning*, 14:115–133, 1994.

[15] A.R. Barron and C.H. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3):1347–1369, 1991.

[16] Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[17] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

[18] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley, 2003.

[19] Peter L. Bartlett and Shahar Mendelson. Empirical Minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

[20] Jean-Patrick Baudry. Clustering through model selection criteria, June 2007. Poster session at One Day Statistical Workshop in Lisieux.

[21] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope Heuristics : Overview and Implementation. Technical Report 7223, INRIA, 2010.

[22] Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[23] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[24] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97:113–150, 1993.

[25] L. Birgé and P. Massart. Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[26] L. Birgé and P. Massart. Gaussian model selection. *J.Eur.Math.Soc.*, 3(3):203–268, 2001.

[27] S. Boucheron and P. Massart. A high dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 2010. To appear.

[28] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[29] P. Burman. Estimation of equifrequency histogram. *Statist. Probab. Lett.*, 56(3):227–238, 2002.

[30] G. Castellan. Modified Akaike's criterion for histogram density estimation. *Technical report ♯99.61, Université de Paris-Sud.*, 1999.

[31] Gwénaëlle Castellan. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8):2052–2060, 2003.

[32] N. N. Čencov. *Statistical decision rules and optimal inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, R.I., 1982. Translation from the Russian edited by Lev J. Leifman.

[33] Michel Crouzeix and Alain L. Mignot. *Analyse numérique des équations différentielles*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1984.

[34] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 1975.

[35] Imre Csiszár and František Matúš. Information projections revisited. *IEEE Trans. Inform. Theory*, 49(6):1474–1490, 2003.

[36] R.M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6(6):899–929, 1978.

[37] R.M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, 1999.

[38] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.

[39] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann.Probab.*, 33:1143–1216, 2006.

[40] Ulf Grenander. *Abstract inference*. New York: Wiley, 1981.

[41] R. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of Probability*, 1:63–87 (electronic), 2005.

[42] T. Klein. Une inégalité de concentration à gauche pour les processus empiriques. *C.R. Acad. Sci. Paris, Ser I*, 334:500–505, 2002.

[43] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5):1902–1914, 2001.

[44] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimisation. *Ann. Statist.*, 34(6):2593–2656, 2006.

[45] Vladimir Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

[46] Vladimir Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3):1332–1359, 2009.

[47] Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):7–57, 2009.

[48] Charles Kooperberg and Charles J. Stone. A study of logspline density estimation. *Comput. Statist. Data Anal.*, 12(3):327–347, 1991.

[49] L.M. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

[50] L.M. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York, 1986.

[51] L.M. Le Cam and G.L. Yang. *Asymptotics in Statistics : Some Basic Concepts.* Springer-Verlag, New York, 1990.

[52] Émilie Lebarbier. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Proces.*, 85:717–736, 2005.

[53] M. Ledoux and M. Talagrand. *Probability in Banach spaces.* Springer, Berlin, 1991.

[54] Vincent Lepez. *Some estimation problems related to oil reserves.* PhD thesis, University Paris-Sud XI, 2002.

[55] Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions, 2009. arXiv:0911.1497v1.

[56] Matthieu Lerasle. Optimal model selection in density estimation, 2009. arXiv:0910.1654.

[57] Matthieu Lerasle. *Rééchantillonnage et sélection de modèles optimale pour l'estimation de la densité de variables indépendantes ou mélangeantes.* PhD thesis, INSA Toulouse, June 2009.

[58] G. Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 1–56. Springer, Vienna, 2002.

[59] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[60] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann.Stat.*, 27:1808–1829, 1999.

[61] P. Massart. *Concentration Inequalities and Model Selection.* Springer-Verlag, 2007.

[62] P. Massart and E. Nédélec. Risks bounds for statistical learning. *Ann.Stat.*, 34(5):2326–2366, 2006.

[63] Cathy Maugis and Bertrand Michel. A nonasymptotic penalized criterion for Gaussian mixture model selection. a variable selection and clustering problems. Technical Report 6549, INRIA, 2008.

[64] D. Pollard. *Convergence of Stochastic Processes.* Springer, New York, 1984.

[65] W. Polonik. Measuring mass concentrations and estimating density contour clusters - an excess mass approach. *Ann.Stat.*, 23(3):855–881, 1995.

[66] Xiatong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *Ann.Stat.*, 22(2):580–615, 1994.

[67] Charles J. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520, Belmont, CA, 1985. Wadsworth.

[68] Charles J. Stone. Uniform error bounds involving logspline models. In *Probability, statistics, and mathematics*, pages 335–355. Academic Press, Boston, MA, 1989.

[69] Charles J. Stone. Large-sample inference for log-spline models. *Ann. Statist.*, 18(2):717–741, 1990.

[70] Charles J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–171, 1994.

[71] Charles J. Stone. Nonparametric M-regression with free knot splines. *J. Statist. Plann. Inference*, 130(1-2):182–206, 2005.

[72] M. Talagrand. On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87 (electronic), 1995/97.

[73] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563, 1996.

[74] M. Talagrand. A new look at independance. *Annals of Probability*, 24:1–34, 1996.

[75] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann.Stat.*, 32:135–166, 2004.

[76] A. B. Tsybakov. *Introduction à l'estimation non-paramétrique.* Springer-Verlag, Berlin, 1996.

[77] S. van de Geer. *Empirical Processes in M-estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[78] Sara van de Geer. Estimating a regression function. *Ann. Statist.*, 18:907–924, 1990.

[79] Sara van de Geer. The method of sieves and minimum contrast estimators. *Mathematical Methods of Statistics*, 4:20–28, 1995.

[80] Sara van de Geer. M-estimation using penalties or sieves. *Journal of Statistical Planning and Inference*, 108(1-2):55 – 69, 2002.

[81] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes.* Springer-Verlag, New York, 1996.

[82] V. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, New York, 1998.

[83] V. Vapnik and A. Chervonenkis. Theory of pattern recognition. *Nauka, Moscow*, 1974.

[84] V. N. Vapnik and A. Ja. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[85] Vladimir Vapnik. *Estimation of dependences based on empirical data.* Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

[86] N. Verzelen. High-dimensional Gaussian model selection on a Gaussian design. Technical Report RR-6616, INRIA, 2008.

[87] Fanny Villers. *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques.* PhD thesis, University Paris-Sud XI, December 2007.

[88] S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, 9(3):60–62, 1938.

[89] Wing Hung Wong and Xiatong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *Ann.Statist.*, 23:339–362, 1995.